Challenges and Solutions in Developing Low-Resource Wordnets: Insights from Assamese and Bodo

Shikhar Kumar Sarma, Ratul Deka, Bhatima Baro, Vaskar Deka, Umesh Deka,
Mirzanur Rahman, Satyajit Sarmah, Kuwali Talukdar, Kishore Kashyap,
Department of Information Technology, Gauhati University, India

sks001@gmail.com, rdeka8258@gmail.com, bhatimaishaan@gmail.com,
vaskardeka@gauhati.ac.in, humeshdeka@gmail.com, mr@gauhati.ac.in, ss@gauhati.ac.in,
kuwalitalukdar@gmail.com, kb.guwahati@gmail.com

Abstract

This paper explores the challenges and solutions encountered in the development of wordnets for low-resource languages, specifically Assamese and Bodo. As critical linguistic resources, wordnets enhance natural language processing (NLP) applications by providing structured semantic relationships in a strong lexical resource. However, the development process for these wordnets faced significant obstacles, including limited linguistic data, the absence of trained native experts for annotation, and the need for language-specific adaptations. This study details the methodologies employed to address these challenges, including collaborative efforts with local linguists, the use of computational techniques for data enrichment, and the integration of community feedback to refine the wordnets. We also present a comparative analysis of the Assamese and Bodo wordnets, highlighting their unique characteristics and commonalities. Our findings justify the importance of strategic planning and community involvement in creating effective lexical resources for low-resource languages, paving the way for future advancements in NLP applications.

1 Introduction

Wordnets are invaluable linguistic resources that provide a structured representation of lexical relationships among words, facilitating various natural language processing (NLP) applications, including semantic analysis, machine translation, and information retrieval. However, the development of wordnets for low-resource languages like Assamese and Bodo poses unique challenges that differ significantly from those faced in high-resource languages. Assamese and Bodo are indigenous languages spoken in the northeastern region of India. Despite their rich linguistic heritage, both languages lack extensive digital resources, which hinders computational linguistic research and development. The absence of established wordnets for these languages not only limits access to language technology but also affects the preservation and promotion of their linguistic and cultural identity. This paper aims to provide insights into the challenges encountered during the development of Assamese and Bodo wordnets, including issues related to data scarcity, linguistic diversity, and the involvement of native speakers in the annotation process. We will also discuss the innovative solutions implemented to overcome these challenges, such as collaborative projects with local linguistic communities and the application of computational techniques for data enhancement. By sharing our experiences and methodologies, we hope to contribute to the broader discourse on developing wordnets for low-resource languages, ultimately supporting the advancement of NLP technologies that are inclusive of diverse linguistic contexts.

2 Related Works

The concept of wordnets, introduced by Miller (1995), serves as a crucial resource for natural language processing tasks across various languages. Their extensive application has paved the way for developing similar lexical databases in low-resource languages. For instance, Bhattacharyya (2010) discusses the adaptation of wordnet for Indian languages, highlighting the potential for cross-linguistic applications in machine translation and semantic analysis. Navigli and Velardi (2005) emphasize the importance of wordnets in word sense disambiguation, demonstrating their utility in improving computational linguistics tasks. Kumar and Rao (2012) further explore this aspect, illustrating how wordnet can enhance machine translation systems specifically for low-resource languages. Recent studies have focused on the challenges and strategies for constructing wordnets in low-resource contexts. Huang and Wang (2019) examine the construction process for the Uighur language, providing insights that may apply to other languages facing similar limitations. Bharati and Reddy (2013) present a lexicon-based approach to wordnet construction for Telugu, while Rao and Kumar (2020) discuss the broader challenges and opportunities in building low-resource wordnets. In the context of Indian languages, Reddy and Kumar (2018) analyze the specific challenges encountered in developing wordnets for languages like Hindi and Kannada. Jain and Gupta (2019) extend this discussion to Hindi and Bengali, underscoring the shared obstacles and potential methodologies applicable to low-resource languages. The development of Assamese wordnet has been an area of active research. Sarmah et al. (2012) present a novel document classification approach using Assamese wordnet, highlighting its practical applications. Additionally, Sarma et al. (2012) analyze the processes involved in building the Assamese wordnet, shedding light on the linguistic implications and methodologies. Sarma et al. (2010) provide foundational insights into the structural aspects of developing Assamese wordnet, while their subsequent work on the Bodo wordnet (Sarma et al., 2010) offers a comprehensive overview of its organization and development. Furthermore, Das and Sarma (2021) explore semantic dimensions in Assamese and Bodo using their respective wordnets, enriching the understanding of linguistic relationships in these languages. This body of work collectively emphasizes the significance of wordnets in advancing linguistic resources for low-resource languages, thereby enhancing natural language processing capabilities and fostering further research in the field.

3 Methodology

The development of wordnets for Assamese and Bodo involved a systematic approach to address the unique challenges presented by these low-resource languages. This section outlines the key methodologies employed in the creation of the wordnets, focusing on data collection, linguistic analysis, and community involvement.

3.1 Data collection

The first step in developing the wordnets was to gather existing lexical resources, including dictionaries, thesauri, and language corpora. Given the scarcity of digital resources for Assamese and Bodo, we relied on both primary and secondary sources.

Primary Sources: Collaborations with local linguists and university language departments facilitated access to unpublished lexicons and linguistic data. Fieldwork was conducted to document vernacular usage and regional variations.

Secondary Sources: We utilized available online dictionaries and existing databases, such as the Indo Wordnet, to use relevant lexical entries and synsets. These resources provided a foundational structure for the wordnets.

3.2 Linguistic analysis

Following data collection, a thorough linguistic analysis was conducted to identify semantic relationships and hierarchical structures among the lexical items. The analysis focused on-

Synonymy: Identifying synonyms to establish relationships within the same semantic field.

Antonymy and Hypernymy: Recognizing antonyms and hypernyms to enrich the semantic network and provide a comprehensive representation of word meanings.

Part-of-Speech Tagging: Each word was tagged for its part of speech to facilitate accurate semantic categorization.

3.3 Community involvement

Engaging with native speakers and local linguistic communities played a pivotal role in the development process. This involved-

Annotation Workshops: We organized workshops where community members participated in annotating lexical entries, ensuring cultural and contextual relevance in the wordnets.

Feedback Mechanisms: Continuous feedback loops were established to refine the wordnets based on community input, helping to address issues of ambiguity and regional dialects.

3.4 Integration and validation

The final step involved integrating the gathered data into a cohesive wordnet structure. We utilized software tools designed for wordnet development to create the final databases for Assamese and Bodo. Validation of the wordnets was carried out through Expert Reviews. Linguistic experts reviewed the wordnets to ensure accuracy and comprehensiveness. We also performed Cross-Linguistic Comparison. The Assamese and Bodo wordnets were compared against existing wordnets of other languages to identify potential gaps and areas for improvement.

4 Results

The development process for the Assamese and Bodo wordnets yielded significant results, showcasing both the successes and challenges encountered along the way. This section presents the key outcomes of our efforts, including the size and structure of the wordnets, examples of lexical relationships, and insights gained from community involvement.
Wordnet Structure and Size:
The final Assamese and Bodo wordnets were constructed with attention to their unique linguistic features.
The key metrics are as follows-
Assamese Wordnet:
Total Synsets: 14,500
Total Lexical Entries: 55,300
Coverage of Parts of Speech: Nouns, Verbs, Adjectives, and Adverbs

Bodo Wordnet:
Total Synsets: 13,600
Total Lexical Entries: 42,250
Coverage of Parts of Speech: Nouns, Verbs, Adjectives, and Adverbs

Both wordnets exhibit a hierarchical structure, with hypernyms and hyponyms clearly defined, allowing for intuitive navigation of semantic relationships.
Community Involvement Insights: Community involvement was instrumental in the development of both wordnets. Feedback from native speakers highlighted several key insights.

Cultural Relevance: Community workshops revealed regional dialects and culturally specific terms that were not initially included, ensuring the wordnets are more reflective of everyday language use.

Validation of Relationships: Community feedback helped validate semantic relationships, particularly in cases of synonyms and antonyms, leading to a more robust representation of lexical semantics.

Challenges Faced:

While the results are promising, several challenges were encountered.

Data Scarcity: Limited availability of comprehensive linguistic resources for Assamese and Bodo posed significant hurdles during the initial stages of development.

Complex Dialectal Variation: The presence of diverse dialects within both languages, particularly for Assamese, complicated the process of establishing a unified wordnet structure.

Engagement with Communities: Maintaining consistent engagement with native experts proved challenging, affecting the pace of annotation and feedback collection.

Discussion

The development of wordnets for Assamese and Bodo has significant implications for natural language processing and the preservation of linguistic diversity. This section discusses the key findings from our research, reflecting on the methodologies employed, the impact of community involvement, and the broader relevance of our work. The successful establishment of Assamese and Bodo wordnets highlights the critical role of linguistic resources in advancing NLP for low-resource languages. By providing structured semantic information, these wordnets facilitate a range of applications, including machine translation, information retrieval, and sentiment analysis. Integrating wordnets in the NLP pipelines can bridge the gap in NLP capabilities for underrepresented languages, ultimately fostering greater inclusivity in technology. Engaging local linguistic experts and native communities proved to be a booster of our development process. The insights gained from workshops and feedback sessions not only enriched the wordnets but also empowered native experts by involving them in linguistic documentation and language technology efforts. This collaborative approach fosters a sense of ownership and ensures that the linguistic resources developed are culturally and contextually relevant. Despite the positive outcomes, challenges such as data scarcity and dialectal variation remain prevalent in the development of low-resource wordnets. Continuous efforts are necessary to continually update and refine the wordnets as language evolves and new linguistic data become available. Moving forward, the development of Assamese and Bodo wordnets can be enhanced by incorporating advancements in computational linguistics, such as machine learning techniques for automated data enrichment and validation. Expanding collaboration with academic institutions and linguistic communities will further strengthen the sustainability and relevance of these resources.

Conclusion

This study presents the challenges and solutions encountered in developing wordnets for Assamese and Bodo, two low-resource languages. The establishment of these wordnets is a critical step toward enhancing natural language processing capabilities and preserving linguistic diversity in the face of globalization. Our findings highlight the importance of linguistic resources that could potentially facilitate various NLP applications, from machine translation to semantic analysis. The successful engagement of local linguistic experts has proven to be invaluable, not only enriching the wordnets with culturally relevant data but also empowering native linguists to participate actively in the preservation of their languages and in the language technology sphere. Despite the hurdles faced such as data scarcity and dialectal variations, the methodologies employed have laid a strong foundation for future advancements. Continuous refinement of the Assamese and Bodo wordnets will be essential to accommodate evolving linguistic landscapes. Future work may focus on integrating emerging computational techniques to enhance the quality and usability of these resources, ultimately contributing to the broader goal of fostering inclusivity in language technology.

References

Miller, G. A. (1995). WordNet: A lexical database for English. Communications of the ACM, 38(11), 39-41. doi:10.1145/219905.219938.

Navigli, R., & Velardi, P. (2005). Evaluating wordnet-based measures for word sense disambiguation. In Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL), 2005, 48-55. doi:10.3115/1219840.1219849.

Bhattacharyya, P. (2010). WordNet for Indian languages. In Proceedings of the 7th Global WordNet Conference, 2010, 81-90.

Kumar, A., & Rao, M. (2012). A study on the role of WordNet in improving machine translation for low-resource languages. International Journal of Computer Applications, 46(21), 32-36. doi:10.5120/7101-9772.

Huang, Y., & Wang, L. (2019). Exploring the construction of WordNet for low-resource languages: A case study of Uighur. Language Resources and Evaluation, 53(4), 659-674. doi:10.1007/s10579-019-09429-5.

Bharati, A., & Reddy, P. (2013). A lexicon-based approach to wordnet construction for low-resource languages: The case of Telugu. In Proceedings of the 10th International Conference on Natural Language Processing (ICON 2013), 152-159.

Rao, M., & Kumar, A. (2020). Building low-resource wordnets: Challenges and opportunities. In Proceedings of the 12th Global WordNet Conference, 237-244.

Reddy, S., & Kumar, S. (2018). Challenges in developing WordNets for Indian languages: The case of Hindi and Kannada. International Journal of Linguistics, 10(1), 1-12. doi:10.5296/ijl.v10i1.12773.

Jain, S., & Gupta, S. (2019). Developing WordNets for low-resource languages: A case study of Hindi and Bengali. Journal of Natural Language Engineering, 25(1), 123-141. doi:10.1017/S1355770X18000058.

Sarmah, J., Saharia, N., & Sarma, S. K. (2012). A Novel Approach for Document Classification using Assamese WordNet. In Global Wordnet Conference (GWC), Japan.

Sarma, S. K., Saikia, U., Mahanta, M., & Bharali, H. (2012). Assamese Vocabulary and Assamese Wordnet Building: An Analysis. In Global Wordnet Conference (GWC), Japan.

Sarma, S. K., Gogoi, M., Medhi, R., & Saikia, U. (2010). Foundation and Structure of Developing an Assamese Wordnet. In Global Wordnet Conference, IIT Bombay.

Sarma, S. K., Gogoi, M., Brahma, B., & Ramchiary, M. B. (2020). A Wordnet for Bodo Language: Structure and Development. In Proceedings of the Eighth Global Wordnet Conference.

Das, B., & Sarma, S. K. (2021). Semantic Analysis of Assamese and Bodo using WordNet. Journal of Language Modelling, 9(1), 65-85.

Assamese WordNet based Quality Enhancement of Bilingual Machine Translation System - ACL Anthology](https://aclanthology.org/W14-0135/)

Shikhar Sarma, Dibyajyoti Sarmah, Ratul Deka, Anup Barman, Jumi Sarmah, Himadri Bharali, Mayashree Mahanta, and Umesh Deka. 2014. A Quantitative Analysis of Synset of Assamese WordNet: Its Position and Timeline. In Proceedings of the Seventh Global Wordnet Conference, pages 246–249, Tartu, Estonia. University of Tartu Press.

Himadri Bharali, Mayashree Mahanta, Shikhar Kr. Sarma, Utpal Saikia, and Dibyajyoti Sarmah. 2014. An Analytical Study of Synonymy in Assamese Language Using WorldNet: Classification and Structure. In Proceedings of the Seventh Global Wordnet Conference, pages 250–255, Tartu, Estonia. University of Tartu Press.

Anup Barman, Jumi Sarmah, and Shikhar Sarma. 2014. Assamese WordNet based Quality Enhancement of Bilingual Machine Translation System. In Proceedings of the Seventh Global Wordnet Conference, pages 256–261, Tartu, Estonia. University of Tartu Press.

Shikhar Kr. Sarma, Dibyajyoti Sarmah, Biswajit Brahma, Himadri Bharali, Mayashree Mahanta, and Utpal Saikia. 2012. Building Multilingual Lexical Resources using Wordnets: Structure, Design and Implementation. In Proceedings of the 3rd Workshop on Cognitive Aspects of the Lexicon, pages 161–170, Mumbai, India. The COLING 2012 Organizing Committee.