Automatic Taxonomy Expansion in IndoWordNet

(With special reference to Assamese WordNet)

Thesis submitted in partial fulfilment of the requirements
for the award of the degree of

Doctor of Philosophy

in

Centre for Linguistic Science and Technology

by

Bornali Phukon

Under the supervision of

Dr. Sanasam Ranbir Singh and Prof. Priyankoo Sarmah

Centre for Linguistic Science and Technology
Indian Institute of Technology Guwahati
Guwahati - 781039 Assam India

May , 2023

Abstract-TH-3092_166155002

Automatic Taxonomy Expansion in IndoWordNet

(With special reference to Assamese WordNet)

Abstract

The task of automatic taxonomy expansion plays a significant role in natural language processing (NLP), as it helps to overcome the issue of low coverage in taxonomies. By effectively performing this task, various NLP applications like information retrieval, text classification, and natural language understanding can achieve better accuracy and efficacy. While numerous studies have explored the challenges of automatic taxonomic expansion, the methods and techniques used in these studies may be less effective for taxonomies like WordNet due to their unique structure and organization.

WordNet is a widely used lexical taxonomy of concepts in a language that comprises not only a hierarchical organization of concepts but also information regarding other semantic relations such as synonymy, meronymy, and troponymy among the concepts, which distinguish it from other taxonomies. The creation of WordNets typically involves manual methods; however, currently, a substantial number of WordNets are generated through the expansion approach, such as those included in Indo-WordNet. Despite its widespread usage, creating WordNet is challenging, with two significant problems being *limited coverage* and *missing relations*. The manual creation process of WordNets can result in limited coverage, while the use of the expansion approach for creating WordNets may result in missing relations between concepts and words. While previous studies have sought to address the issue of limited coverage, the problem of missing relations has yet to receive adequate attention. Furthermore, while automatic taxonomy expansion approaches have been proposed to resolve the issue of limited coverage, their effectiveness for WordNet expansion remains in question. The primary reason is that the expansion of WordNet requires not only inserting new concepts (*attach operation*) but also extending existing ones (*merge operation*) as shown in figure 1.4. However, most existing studies on taxonomy expansion only focus on the (*attach operation*). Furthermore, WordNet taxonomies, especially those in Indian languages, tend to have a multi-root structure. It makes it more challenging to utilize traditional methods for the expansion of WordNet taxonomy as these methods are not designed to handle the challenges of a multi-root structure, which may limit their usefulness in expanding WordNet taxonomies. In light of these challenges, this thesis work aims to address the problem of automatic taxonomy expansion by addressing the challenges in WordNet, especially in IndoWordNet. The objective is to develop a solution that can be extended to other taxonomies beyond WordNet.

This thesis first studies the problem of missing synonymy relations in WordNet taxonomy. It considers Assamese Wordnet as a case study. It investigates the effectiveness of Link prediction methods. As WordNets can be visualized as a network of unique words connected by synonymy relations, link prediction in complex network analysis is an effective way of predicting missing relations in a network. Hence, in order to predict the missing synonyms in the Assamese WordNet, link prediction methods were used in the current work that proved effective. It is also observed that for discovering missing relations in the Assamese WordNet, simple local proximity-based methods might be more effective as compared to global and complex supervised models using network

embedding. Second, a novel multi-task learning-based deep learning method known as *Taxonomy Expansion with Attach and Merge (TEAM)* is proposed, which performs both the *merge* and *attach* operations. This is the first study that integrates both the*merge* and *attach* operations in a single model to the best of our knowledge. The proposed models have been evaluated on three separate WordNet taxonomies, viz., Assamese, Bangla, and Hindi. From the various experimental setups, it is shown that TEAM outperforms its state-of-the-art counterparts for *attach* operation and also provides highly encouraging performance for the *merge* operation. Third, As TEAM considers local context, it faces challenges when it is applied to multi-root taxonomies. To address the limitations in TEAM, this thesis proposes another approach, LG-TEAM, which combines both the local and global context of taxonomy in an integrated *attach-merge* expansion environment, providing a more robust solution to the problem of taxonomy expansion. Extensive experiments on English, Assamese, Bengali, and Hindi WordNets demonstrate both the effectiveness and the efficiency of LG-TEAM for automatic taxonomy expansion.