INDIAN INSTITUTE OF TECHNOLOGY GUWAHATI
PhD-17 SHORT ABSTRACT OF THESIS

Name of the Student	:	*PRACHURYYA KAUSHIK*
Roll Number	:	*206201002*
Programme of Study	:	*M.Tech + Ph.D.*
Thesis Title	:	*Multilingual Fine-grained Named Entity Recognition*
Name of Thesis Supervisor(s)	:	*Prof. Ashish Anand*
Thesis Submitted to the Academic Division	:	*Computer Science and Engineering*
Date of completion of Thesis Viva-Voce Exam	:	*14.05.2026*
Key words for description of Thesis Work	:	*Named entity recognition, Fine-grained Named entity recognition, Cross-lingual transfer, Multilingualism, Endangered languages, Resources for less-resourced languages, Automatic creation and evaluation of language resources, Datasets for low resource languages*

SHORT ABSTRACT

The extraction of domain-specific entity information, particularly via Fine-grained Named Entity Recognition (FgNER), is crucial for advancing Natural Language Processing (NLP) applications such as knowledge base construction and relation extraction. However, the creation of high-quality FgNER datasets is severely hampered by the prohibitive cost of manual annotation and the inherent noise in dataset generated through weak supervision paradigms such as distant supervision. This is even more challenging for languages with a diverse landscape and a scarcity of resources. This thesis addresses these critical quality and resource scarcity challenges by developing novel, scalable frameworks for generating vast, high-quality FgNER datasets across numerous languages and diverse entity type taxonomies. In addressing the challenges, this work makes four key contributions. In the first contribution, a Taxonomy Adaptable Fine-grained Entity Recognition through Distant Supervision framework, TAFSIL, is proposed. TAFSIL leverages the high interlink between WikiData and Wikipedia through multi-stage heuristics, fuzzy matching, and quality sentence selection. The proposed framework is then used to create robust FgNER datasets in six Indian languages: Hindi, Marathi, Sanskrit, Tamil, Telugu, and Urdu. With approximately three million samples, the TAFSIL dataset achieves an 83% relative improvement in average F1 score over zero-shot baselines. Although TAFSIL is a highly scalable framework across multiple languages and taxonomies, its dependence on the availability of Wikipedia is a bottleneck. Since Wikipedia is not available for many low-resource and vulnerable languages, this thesis tries to exploit alternative resources and methodologies. The second contribution is an application of the annotation projection method to create FgNER resources for three vulnerable languages of India's North Eastern Region: Bodo, Manipuri, and Mizo. Existing FgNER resources in English and parallel corpora from English to target vulnerable languages are utilized via the annotation projection method to create these high-quality datasets. Moreover, in this work, cross-lingual zero-shot FgNER analyses are discussed substantially. These analyses suggest the importance of script similarity in FgNER, which paves the way to leverage this property to reduce noise and improve dataset quality. The third contribution is CLASSER, a framework for Cross-lingual Annotation Projection enhancement through Script Similarity for Fine-grained Named Entity Recognition. CLASSER employs a two-stage process: an initial projection from source language annotations to target language annotations, followed by a crucial refinement step that leverages datasets from script-

similar languages. CLASSER is applied to five low-resource Indian languages: Assamese, Bodo, Marathi, Nepali, and Sanskrit. The dataset created through the CLASSER framework is of very high quality, demonstrating substantial performance gains, with F1 score improvements of up to 26% in Marathi and 46% in Sanskrit over the current state-of-the-art. Although the CLASSER framework is very effective across script-similar languages, its scalability is limited for languages written using different scripts. Therefore, the fourth contribution consolidates the resource creation efforts by introducing the highly scalable framework, Entity-Anchored Machine Translation (EaMaTa) framework. The EaMaTa framework is then used to create SampurNER, a large-scale, high-quality FgNER dataset covering all 22 scheduled Indian languages, and its extension FiNE-MiBBiC, an FgNER dataset targeting four extremely low-resource languages. Rigorous analyses confirm the superior quality of these resources, demonstrating an up to 9% increase in F1-score over the current state-of-the-art. Extensive insights are highlighted regarding the influence of language family and script similarity on cross-lingual FgNER model performance across one of the most diverse linguistic groups. The collective work presented in this thesis significantly mitigates the resource gap for FgNER in Indian and low-resource languages, providing high-quality, large-scale datasets and establishing effective, scalable frameworks for future resource creation efforts.