INDIAN INSTITUTE OF TECHNOLOGY GUWAHATI
SHORT ABSTRACT OF THESIS

Name of the Student	:	**ABHISHEK**
Roll Number	:	**146201004**
Programme of Study	:	**M.Tech + Ph.D. Dual Degree**
Thesis Title:		**Fine-grained Entity Detection and Typing**
Name of Thesis Supervisor(s)	:	**Dr. Amit Awekar and Dr. Ashish Anand**
Thesis Submitted to the Department/ Center	:	**Computer Science and Engineering**
Date of completion of Thesis Viva-Voce Exam	:	**07/07/2020**
Key words for description of Thesis Work	:	**Natural Language Processing, Entity Detection, Entity Typing**

SHORT ABSTRACT

Detection and typing of entity mentions present in natural language text are one of the fundamental problems in information extraction paradigm. The dissertation directly focuses on advancing the state-of-the-art of the entity detection and entity classification problems in a setting where entity mentions can belong to a large set of types spanning diverse domains such as biomedical, finance, and sports. Moreover, the entity mentions could be mentioned in several text genres, such as newswire, scientific abstracts, and forums. When the scope of entity mentions is diverse, and several text genres are involved, data scarcity becomes one of the primary issues for these tasks.

The thesis addresses several issues related to the data scarcity, either directly or indirectly. First, we propose a noise-aware learning model for the task of fine-grained entity typing. The proposed model outperforms previous state-of-the-art models, which assumes that the training dataset is noise-free. The noise-aware model addresses the data-scarcity issue indirectly as the majority of datasets for the fine-grained entity typing task are generated automatically using the distant supervision paradigm. The automatically generated datasets have noise, and thus noise-aware models permit efficient and effective learning. We also propose transfer learning approaches in cases where the training dataset size is small.

Second, we propose a collective learning framework for the task of fine-grained entity typing. The proposed framework aggregates different datasets which can have partial overlapping labels and can predict a unique fine-grained label for a given entity mention. The work also addresses the data scarcity issue as often we do not have datasets available with all of the label set annotated, and utilizing different datasets that have partial labels annotated eliminates the need to create new datasets.

Third, we propose a framework to improve the quality of the datasets generated in the distant supervision paradigm for the fine-grained entity detection and fine-grained entity typing task. Using the framework, we created two datasets, each containing more than thirty million sentences annotated with around hundred and thousand entity types, respectively. The work directly addresses the data scarcity issue by sharing new datasets for these tasks with the research community.