INDIAN INSTITUTE OF TECHNOLOGY GUWAHATI
SHORT ABSTRACT OF THESIS

Name of the Student	:	LENIN LAITONJAM
Roll Number	:	166101018
Programme of Study	:	Ph.D.

Thesis Title: **Manipuri-English Machine Translation using Comparable Corpus (An Unsupervised Statistical Machine Translation Approach)**

Name of Thesis Supervisor(s)	:	DR. SANASAM RANBIR SINGH
Thesis Submitted to the Department/ Center	:	CSE
Date of completion of Thesis Viva-Voce Exam	:	20 MAY 2022
Key words for description of Thesis Work	:	MACHINE TRANSLATION

SHORT ABSTRACT

Machine Translation (MT) is an essential tool for communicating with foreign-language speakers. The current mainstream MT frameworks, namely, Statistical MT (SMT) and Neural MT (NMT), are characterized by learning to translate automatically via machine learning techniques. It has been observed that these systems require a large number of parallel sentences between the source and the target language pair to produce a high-quality translation. Unfortunately, readily available parallel sentences are limited for most language pairs. Manually generating a quality parallel corpus is also very costly and time-consuming. As a result, many practical applications of MT are restricted to widely spoken and rich-resource languages. On the other hand, MT quality has not reached a reasonable level in many low-resource language pairs.

This thesis reports the problem of developing an MT system that translates between low-resource Manipuri and English. Manipuri is one of the scheduled Indian languages. The study focus on improving the MT quality between the language pair by exploiting unsupervised MT approaches to cope with bilingual corpora's scarceness. Unsupervised MT enables translation between languages without using parallel data by exploiting source and target language monolingual corpora. This thesis first presents a Manipuri-English comparable corpus to facilitate MT research between the language pair. The corpus belongs to the same domain and is also aligned at date and document levels. Although the results are promising, unsupervised MT techniques have the drawback that their performance suffers when the source and target languages have different linguistic properties. To alleviate issues incurred due to different linguistic aspects between English and Manipuri, this thesis proposes two methods. The first method is proposed to normalize the morphological inflection issue of Manipuri. The second method is proposed to induce inter-language connecting points between Manipuri and English.

The last part of the thesis is dedicated to making the best use of the proposed comparable corpus for the language pair MT task. Specifically, the study exploited the document-aligned and temporally-aligned characteristics of the corpus. Firstly, this thesis proposes a multi-step approach to exploit document-aligned comparable corpus. From various experimental results on English-to-Manipuri and Manipuri-to-English MT, it is observed that both the proposed methods developed for leveraging the comparable corpus's different alignment characteristics succeeded in their respective task and further enhanced the translation results.