National Institute of Technology Rourkela

Defence Seminar

Seminar Title	: Exploring Efficacy of Machine Translation System for Indian Languages
Speaker	: Sudhansu Bala Das (Rollno : 520cs6006)
Supervisor	: Dr. Tapas Kumar Mishra
Venue	: Convention Hall, CS Department
Date and Time	: 14 Dec 2024 (11:30 AM)

Abstract

: In era of digital globalization, the necessity of efficient intercommunication among people from diverse language backgrounds is growing exponentially. **Machine Translation System** (MTS) can be utilized to facilitate seamless communication among diverse linguistic communities. In general, Machine Translation (MT) is the process of translating text from one language to another without the need for human intervention. In consideration with the Indian environment, the development of a quality MTS for the Indian Languages (ILs) is in huge demand and a challenging task, since many ILs are treated as low-resource languages. As a result, the performance of MTS built upon these Indian Languages is not upto the mark. However, the vast linguistic diversity and socioeconomic significance of ILs, in India and abroad, serve as the driving forces for a quick attention on this domain. The variability in morphology, syntax, and semantic expression among ILs poses significant obstacles in the process of developing an effective MTS. Addressing such complexities, this thesis presents five important developments aimed at improving the quality and efficiency of MT systems for ILs. **The first contribution presents an effective Statistical Machine Translation (SMT) system designed especially to translate text from English into eleven (11) Indian languages and vice-versa.** To enhance translation quality and clean up the data, many data preprocessing techniques are utilized. The effects of distance-based reordering and Morpho-syntactic Descriptor Bidirectional Finite-State Encoder (msd-bidirectional-fe) reordering on ILs are analyzed. However, fluency and context were problems for SMT, which prompted the development of NMT models for ILs. **Neural Machine Translation (NMT) system that can translate between English and eleven (11) Indian languages in both directions is developed in the second contribution**. The Backtranslation (BT) is utilized for data augmentation to enlarge the dataset. Therefore, the influence of data augmentation on NMT for ILs is investigated to evaluate its efficacy in enhancing translation robustness and quality. Despite this, NMT systems are limited in their ability to translate low-resource languages since learning meaningful cross-language mappings requires enormous amounts of data. Hence, **the third contribution investigates methods for translating from English into eleven Indian languages and vice versa using Multilingual Neural Machine Translation (MNMT).** The MNMT is a technique for MT that builds a single model for multiple languages. It is preferred over other approaches, since it decreases training time and improves translation in low-resource contexts. To enhance translation quality in linguistically diverse contexts, MNMT models are integrated withseveral techniques, including pivot-based machine translation (MT), backtranslation, and language-relatedness. **The fourth contribution expands the capability of MNMT to include intra-Indian language pairs.** The effect of the grouping of related languages, namely, East Indo-Aryan (EI), Dravidian (DR), and West Indo-Aryan (WI) on the MNMT model are examined. The role of pivot-based MNMT models in enhancing translation quality is investigated. Owing to the presence of large good-quality corpora from English (EN) to ILs, MNMT IL-IL models using EN as a pivot are built and examined. Furthermore, the effect of transliteration on ILs is also analyzed. To explore transliteration, the best MNMT models from the previous approaches (in most of cases pivot model using related groups) are determined and built on corpus transliterated from the corresponding scripts to a modified Indian language Transliteration script (ITRANS). **The final contribution assesses how SMT, NMT, and MNMT systems for Indian languages are affected by subword tokenization techniques like SentencePiece, WordPiece, and Byte-Pair Encoding (BPE).** These methods are essential for improving translation performance and managing morphologically complex languages. All the approaches vis-a-vis their contributions are evaluated using standard evaluation metrics.