National Institute of Technology Rourkela

Synopsis Seminar

Seminar Title	: Exploring the efficacy of Machine Translation systems for Indic Languages
Speaker	: Sudhansu Bala Das (Rollno : 520cs6006)
Supervisor	: Dr. Tapas Kumar Mishra
Venue	: Convention Hall, CS Department
Date and Time	: 27 Jun 2024 (17:00)
Abstract	: In an era among where globalization is mostly driven by digital material, the importance of efficient interaction Indian languages (ILs) is growing. Effective machine translation (MT) systems are critical to facilitate seamless communication and remove obstacles between diverse linguistic communities. 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 machine translation system (MTS) for the Indian languages (ILs) is in huge demand. Still, it remains a challenging task, since many ILs have low in terms of resources, resulting in an adverse impact on their translation quality. The vast linguistic diversity and socioeconomic significance of Indian languages both within the Indian Union and internationally serve as the driving forces for the attention on these languages. The several linguistic groups that make up Indian languages, such as Dravidian, Indo-European, Indo-Aryan, and others, are distinguished by their distinct syntactic patterns and cultural subtleties. The variability in morphology, syntax, and semantic expression poses significant obstacles to machine translation. Hence, this thesis presents five important developments aimed at improving the quality and efficiency of machine translation (MT) systems for ILs. To start, the first contribution presents an effective Statistical Machine Translation (SMT) system designed especially to translate text from English into 11 Indian languages and viceversa. To enhance translation quality and clean up the data, many data filtration techniques are studied. The effects of distance-based reordering and Morpho-syntactic Descriptor Bidirectional Finite-State Encoder (msd-bidirectional-fe) reordering on ILs are analyzed. Following this work, the second contribution describes a Neural Machine Translation (NMT) system that can translate between English and 11 Indian languages in both directions. Backtranslation (BT) is utilized for data augmentation to enlarge the dataset. BT has advantages to many languages. However, whether it has a significant impact on ILs remains debatable. Hence, the influence of data augmentation on neural machine translation (NMT) for Indian languages is investigated to evaluate its efficacy in enhancing translation robustness and quality. However, NMT systems are restricted in translating low-resource languages as a huge quantity of data is required to learn useful mappings across languages. Hence, the third contribution investigates methods for translating from English into eleven Indian languages using Multilingual Neural Machine Translation (MNMT) and vice versa. Multilingual neural machine translation (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, i.e., for languages that have insufficient corpus. The integration of techniques like pivot-based machine translation (MT), backtranslation, and language-relatedness improves translation accuracy and fluency in linguistically varied environments. The fourth contribution expands MNMT&rsquos capability 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 in this contribution. 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 and each contribution are evaluated using standard evaluation metrics.