Researchers devise digital method to process Sanskrit texts

New Delhi, March 25 (India Science Wire): While various digital resources have improved the accessibility and use of world languages and even regional languages, Sanskrit presents unique challenges in automated computational processing. In addition to the sheer volume and diversity, both stylistic and chronological, found in these texts, the linguistic peculiarities expressed by the language, pose several challenges in making these works accessible to the world. Researchers at the Indian Institute of Technology (IIT) Kharagpur are making Sanskrit accessible with their Artificial Intelligence (AI)-based system for processing Sanskrit texts.

Researchers led by Dr Pawan Goyal have developed a digital infrastructure for the efficient processing of Sanskrit texts, by effectively combining state-of-the-art machine learning techniques and traditional linguistic knowledge from Sanskrit. The proposed framework is based on Energy-based models and it enables the encoding of relevant linguistic information as constraints.

“Processing of Sanskrit texts poses several challenges owing to the high lexical productivity of the words, free word order in poetry, euphonic assimilation of sounds at the word boundaries and phonemic orthography followed in writing. Keeping these in mind, we proposed a generic graph-based framework that takes advantage of the free word-order nature of the language. Further, we make use of linguistic insights from the traditional Sanskrit grammar for learning the feature function and applying the relevant constraints.” explained Dr. Goyal.

“Our proposed framework substantially reduces the training data requirements to as low as 10%, as compared to that of the neural state-of-the-art models. In all the Sanskrit-related tasks discussed in the work, we either achieve state-of-the-art results or ours is the only data-driven solution for those tasks, added Dr Goyal.”

This work is accepted for publication in the Computational Linguistics journal published by the MIT Press. This work has been carried by research scholar Dr. Amrith Krishna, currently, a post-doctoral fellow at the University of Cambridge, supervised by Dr. Pawan Goyal. The paper currently addresses the tasks of word segmentation, morphological parsing, dependency parsing and poetry to prose conversion of Sanskrit text.

The classical language has a rich literary tradition spanning more than two millennia that encapsulates the cultural ethos of this civilizational nation. Works in Sanskrit, numbering more

than 30 million extant manuscripts, include extensive epics, subtle and intricate philosophical, mathematical, and scientific treatises, and rich literary, poetic, and dramatic texts.

The proposed AI-based system, used in conjunction with interactive tools such as the Sanskrit Heritage reader, can aid the users in the easier analysis of these manuscripts with word-by-word analysis and translation, the relation between words, poetry to prose conversion, search and question answering, etc.

The research team is now actively collaborating with several external research groups to extend the application of the proposed system for automatic speech recognition and question-answering in Sanskrit. (India Science Wire)

Keywords: Digital resources, languages, regional languages, Sanskrit, Computational processing, linguistics, Indian Institute of Technology, IIT Kharagpur, Artificial Intelligence, AI

ISW/USM/ENG/25/03/2021