Natural Language Processing (NLP): A means for human-computer interaction using the Khasi Language

"Tham Annotated Khasi Corpus"

By Dr Medari Janai Tham*

Natural language processing or NLP enables a computer to analyse and understand human languages in written or spoken form. It is the backbone in the development of machine translation software, voice assistants such as Google Assistant, Siri, chatbots, automatic spelling correction, grammar checking, search engines, and so on.

Machine learning is one of the techniques employed by NLP to carry out the above-mentioned tasks. As the name suggests, machine learning implies teaching a machine to learn to execute desired tasks by providing one essential component for learning which is data. This data is formatted according to the task at hand and provide a way for a machine learning model to be generated.

This model is a mathematical representation and it is responsible for evaluating new cases which are similar to the data it has seen. Hence, the development of such data is an important step for any NLP task. In NLP literature, this data is called *corpus* which is a collection of machine-readable text that is sampled to be representative of a particular language. Such corpora exist for languages such as English, German, Chinese, Hindi, Bengali, Punjabi, etc. In English, the most widely used corpora are the British National Corpus (BNC) and it is popular among researchers due to its accessibility. However, not all of these corpora are easily accessible.

Khasi Corpus Construction

For resource-poor languages like Khasi, an Austro-Asiatic language family spoken mainly in the state of Meghalaya and some parts of Assam and Bangladesh, there is no available corpus till recently. However, now the Khasi annotated corpus titled "**Tham Khasi annotated corpus**" has been developed and is freely accessible through the European Language Resources Association (ELRA) via the link http://catalog.elra.info/en-us/repository/browse/ELRA-W0321/. The corpus comprises Khasi sentences formatted with parts-of-speech tags using the formulated BIS (Bureau of Indian Standards) POS (Parts-of-Speech) tagset to ensure standardised tagging with other Indian languages. This corpus can serve as a resource for researchers to perform various NLP tasks for the Khasi language.

One such outcome is NLP tools for Khasi currently available in https://grammarkhasi.in comprising of POS taggers that automatically tag any given Khasi sentence with its parts of speech and a shallow parser that automatically detects Khasi noun phrases and verb phrases in a sentence. Corpus analysis has revealed the existence of 560 words that are multi-functional which lexicographers or any interested researchers can further investigate. The Tham Khasi annotated corpus is part of the PhD research output of Dr Medari Janai Tham, Associate Professor, Department of Computer Science, St. Anthony's College, Shillong.

Conclusion

Developing language technology tools for an under-resourced language such as Khasi has been challenging and simultaneously exhilarating to discover the nitty-gritty of the language in the way studies such as this one exposes. The performance of the HMM tagger conditioned with the features intrinsic in the language has shown that it also provides good performance as reported in the literature relating to HMM POS taggers. This work, being a new initiative, annotating the corpus and developing the tagger, is limited by available resources; however, increasing the size of the annotated corpus for further analysis will be a good step forward.

******

**The writer is PhD Research Scholar, Department of Computer Science, St. Anthony's College, Shillong.)*