

Manipuri NLP Workshop Report (Department of Computer Science, Manipur University, Canchipur), 3rd – 13th February, 2012

The Orientation cum Training Program on Natural Language Processing was organized by LDC-IL in collaboration with the Department of Computer Science, Manipur University, Canchipur, Imphal. The program was well thought-out from the 3rd to 13th February, 2012, at the Lecture Hall (Second Floor) of the Department of Computer Science. There were 55 participants, excluding the Department staff and helpers, out of which 30 were trainees selected for the program through screening, 15, the Resource Persons engaged for presenting the talks and 10, novices in the field of the entire orientation program.

The first day started off with an inaugural function. Prof. H. Nandakumar Sarma, Vice-Chancellor, Manipur University (MU), Dr. Tejkumar Sinam, Head (i/c), Department of Computer Science, MU, Prof. M. Dhaneshwar Singh, Dean (i/c), School of Mathematical & Physical Science, MU, kindly consented to grace the function as the Chief Guest, President and Guest of Honour respectively and Mr. A. Nandaraj Meetei, Co-ordinator, Lecturer/RP, LDC-IL, and Dr. H. Mamata Devi, Local co-ordinator, Assoc. Professor, MU took their respective chairs. Prof. H. Nandakumar Sarma kindled the inaugural light and delivered his speech on the atomic characteristics of the elements within human language in a metaphoric sense of physical world. Dr. Tejamani Sinam spoke about the historical development of NLP and tools and activities undertaken by the researchers. Prof. M. Dhaneshwar Singh talked about the academic encouragement in terms of NLP, particularizing the tonal feature of Manipuri Language. Mr. A. Nandaraj Meetei talked about the objectives of LDC-IL with respect to the repository of linguistic resources in all Indian languages, facilitation of database creation and training through workshops, seminars etc. He also instructed the selected candidates as well as all the participants to attend all the days of the program and asked them to have interactive sessions. Dr. H. Mamata Devi gave her speech on the goals of the workshop stating that the targets of orientation cum training program are to disseminate the knowledge of NLP amongst student community, to equip students to work or pursue research in language technology and to promote technology development in Indian Languages. She mentioned some of the NLP tools developed by the department and also asked the participants to have a look on such undertaking. Dr. N. Gourakishwar, Senior lecture, Department of Computer Science, MU, gave vote of thanks for the inaugural session by opening the fresh eyelids of every participant in the hall with his quick-to-remind delivery.

For the entire orientation cum training program, after the Inaugural Function, Dr. N. Pramodini Devi, Assoc. Professor, Linguistic Department, MU, first gave a class on Phonetics. She well illustrated the general production of speech sound in terms of place as well as manner of articulation. Some typical Manipuri phonemes were taken into consideration during her class and participants from non-linguistics had been made perceive the existing phenomena of human articulation system. Prof. P. Madhubala Devi, Head, Linguistic Department, MU gave a class on *Phonology*. Within the phonological system of Manipuri, she explained the distinctive features of Manipuri phonemes. Underlying correlations between phonemes and allophones, particularly in Manipuri, were also exemplified thereafter.

On the second day Dr. H. Mamta Devi, MU, presented a paper on *Artificial Intelligence (AI) & NLP/CL*. She basically talked about how AI should be fitted in the Computer Science taxonomy. Dr. I. Gambhir Singh, Reader, English Department, MU, presented a paper on *Natural Languages processing and its utility in information technology*. The nature of the presentation was so innovative that participants paid attention to a great extent to the up-to-date information contained in it. Dr. S. Imoba Singh, Assoc. Professor, MU, gave a special lecture on *Lexicography*. He emphasized on the semantically related lexical items that are to be made entries in the dictionary by tracing back the diachronic changes of the lexemes of Manipuri in particular.

On the third day Dr. I. Robindro Singh, lecturer, NERLC, Guwahati, presented a paper on *Manipuri Morphology*. He described both inflectional and derivational morphemes and their various categorial levels reflected within their occurring domains. Subsequently, Prof. Ch. Yashawanta Singh, Linguistic Department, MU delivered a presentation on *Basic Syntax of Manipuri*. He mainly focused on the different types of sentence on the basis of structural and functional points of view. As a second lecture, Dr. S. Imoba Singh gave a lecture on *Semantics*. He narrated the theories of meaning and componential analysis of linguistic features which an item carries. As a final presentation of the same day, Mr. A. Nandaraj Meetei, LDC-IL, CIIL, Mysore gave a special lecture on *Morpho-syntactic features-Accounting for Manipuri Morphological Analyzer*. The key point of the lecture was to display the morpho-syntactic features both for nominal and verbal domains to be incorporated into the Morphological analyzer of Manipuri.

On the fourth day Mr. L. Anand Singh, Senior Research Assistance, LDC-IL, CIIL, Mysore presented a paper on *Corpus - Concept, Types, Balance and Tools*. He spoke about the origin of corpus development and different types of corpus and related tools explaining the criteria of a balanced corpus that covers a wide range of text categories representative of the language or language variety under consideration. Next to it, Dr. M. Bidyarani Devi, Junior Resource Person, LDC-IL, CIIL, Mysore gave a presentation on *Linguistic Knowledge and Corpus- Cleaning the text*. She dictated the necessity of cleaning the raw data appeared in the texts before making use of the data through computational application. And, she further added that the linguistically cleaned data can be successfully used as training data for many NLP applications. As a last presentation for the same day, Miss Y. Premila Chanu, Junior Resource Person, LDC-IL, CIIL Mysore presented a paper on *Speech Segmentation & Annotation*. Primarily, she mentioned about the speech corpus and their uses. In succession, she came to explain how speech data had to be collected with the help of Speech guidelines from different parts of the state of Manipur in particular. In the later part of her paper, there were some hands on experience on annotation of speech data. The fourth day was ended with Tool Demo and Practice Session.

On the fifth day Mr. A. Nandaraj Meetei, LDC-IL, CIIL, Mysore sequentially presented papers on (i) *POS system: Categories and sub-categories in Manipuri* (ii) *POS annotation: Hierarchical tagging and LDC-IL tools* and (iii) *Verbal Noun as Verb* respectively. To be precise, the first paper analyzed the feasible major part-of-speech categories and sub-categories of Manipuri POS system. The second paper described the hierarchical tagging incorporated in the POS LDC-IL tools showing hands on experience on annotation using LDC-IL annotation tool of version-3. The third paper was about the so-called Verbal Noun, a mixed category in the POS system, where a verbal root becomes a nominal after affixing a suffix called nominalizer. The paper, providing various typical characteristics, made an attempt to argue that such a category should be placed under the verbal categories as it behaves like a verb from the morpho-syntactic perspectives, too.

On the sixth day Mr. L. Anand Singh, LDC-IL, CIIL, Mysore gave a presentation on *Corpus-based Lexicography*. It was insightful that, by consulting a corpus containing a rich amount of textual information, a lexicographer can become more confident over the obtaining results which reflect the actual meaning of a particular word more accurately. In fact, a corpus-based dictionary can be revised much more quickly than the manual one, providing the up-to-date information about language. Mr. Rajesha N, Lecturer/RP, LDC-IL, CIIL, Mysore presented a paper on *Encoding: Concept of Font and Unicode*. Basically, the paper was focused on representing texts in computer systems, and standardizing Information Interchange. Standardization of Encoding includes ASSCI, ISSCI, and UNICODE, etc. After that, he also presented special lecture on *Transliteration- BANGALA to MEETEI MAYEK*, where a demo tool was shown to illustrate how the two scripts were interfaced.

On the seventh day, as a second presentation, Dr. I. Robindro, NERL, gave lecture on *Linguistic Theories and Computer Application*. Dr. Robindro put emphasis on implementation of linguistic theories into the domain of computer application, which is the core of artificial intelligence. Also as a second lecture, Dr. Th. Mamata

Devi, MU, gave a presentation on Corpus Utilities and Corpus analysis tools: Transliteration, Frequency, N-gram, KWIC-KWOC, Concordances, Extraction. The presentation was well informative and interactive that two project staff demonstrated already developed tools. As a last presentation of the same day, Dr. Kh. Aruna Devi, Assoc. professor, Department of English, Oriental College, Imphal, one of the trainees in the program, gave a presentation on The role of English in Language Teaching. The paper discussed the some of the basic phonological systems of Manipuri and English and made attempt to lay down the underlying parameters between the two languages concerned such as- English phonemes- the initial voiced fricative, /v/, the initial aspirated fricative /f/; the final voiced /b/, /d/ /g/ etc. and their counter parts in Manipuri, where they are not properly spelled out. Dr. Aruna showed her full interest in the program and initiated presenting a self-contained paper amongst the trainees.

On the eighth day, Mr. Rajesha N, LDC-IL, CIIL Mysore, presented a paper on Machine Learning: Statistical Approaches. Basics of the class were focused on introduction to the two major approaches of MT, namely, statistical approach and rule based approach. The class was extended even to the domain of WSD (Word Sense Disambiguation). Dr. Utpal Sharma, Reader, Department of Computer Science and Engineering, Tezpur University, gave two lectures consecutively on *Natural Language Processing and Artificial Intelligence and Approaches to NLP: Computational Grammar Approach and Data driven/Inductive Approach*. He Dr. Utpal also discussed some insightful areas of Decision Trees and Decision Lists and HMM. He illustrated the basic neural network and genetic algorithm in an interactive trend.

On the ninth day, as a second lecture, Prof. Ch. Yashawanta Singh, Linguistic Department, Manipur University, presented a paper on *WordNet: Concept and Applications, Development & Challenges*. Professor Yashawanta focused on the lexical database for the Manipuri language grouping the language concerned into sets of synonyms called synsets. The paper provided general definitions and highlighted various semantic relations between the synonym sets. Dr. R.K. Musuksana Devi, Reader, Manipuri Department, D.M. College of Arts, Imphal, one of trainees in the program, also gave a presentation on *Some Elements of Manipuri Grammar*. The paper discussed basic morphological and some phonological properties of the language concerned. Dr. Musuksana, showing enthusiastic interest in the program, dictated all the contents of the paper. In a brief manner, Mr. Rajesha N, LDC-IL, CIIL, Mysore, gave a class on OCR: *Concept and Application*. The paper was introductory that it tells about how OCR systems for Indian languages had been brought in picture, beginning with a birth story of an *Apsara* named *Urvashi* from Hindu Mythology. Trainees enjoyed the story as well as its intellectual connection to the introduction to OCR. Dr. Th. Tangkeshwar Singh, Assist. Professor, Department of Computer Science, MU, gave a presentation on *Hand Written Character Recognition*. Shree Tangkeshwar's paper was so innovative that the trainees could get access to the logical insight of Meetei/ Meitei scripts and their existing figures and the diachronically related material finding and preservation. Mr. S. Somorjeet Singh, Guest Lecturer, Department of Computer Science, MU, gave a presentation, as a last paper of the same day, on *Data Structure for E-dictionary*. He illustrated the data structure to be incorporated into the making of e-dictionary in terms of the organization of head word, pronunciation, meaning in a systematic procedure.

On the tenth day, the last day of the program, Mr. A. Nandaraj Meetei, LDC-IL, CIIL, Mysore, gave a presentation on *Parsing and Chunking: Basics of Dependency Relation in Manipuri*. The core idea of the paper was to demonstrate that Manipuri explores dependency relation between the morpho-syntactic heads and their corresponding complements in a functional sequence of head-complement relation. After the paper presentation, there was a short program on a formal Valedictory Talk. The valedictory function was chaired by Dr. O. Imocha Singh, Assoc. Professor, Head, Department of Computer Science, MU. He delivered speech encouraging the spirit of the organizing program committee. Mr. A. Nandaraj Meetei, LDC-IL, Co-ordinator of the program, extended the thanks-giving speech. Dr. H. Mamata Devi, MU, Local-coordinator of the program extended her heart-felt congratulation for the complete orientation cum training course for the entire ten

days. For the comments and suggestions for more improving the programs to follow were from amongst the trainees and presenters. Dr. N. Gourakishwar Singh, Senior lecture, Department of Computer Science, MU, who gave vote of thanks for the inaugural session, also gave vote of thanks for closing session by reminding everyone there of the fully academic vibrant past days in the conference hall. The chair person of the closing day declared the *Ten Day Orientation cum Training Program on Natural Language Processing closed*.