ICON 2021

The 18th International Conference on Natural Language Processing

Proceedings of the First Workshop on Parsing and its Applications for Indian Languages (PAIL)

December 16, 2021

Preface

A workshop on Parsing and its Applications for Indian Languages (PAIL-2021) was organized in conjunction with 18^{th} International Conference on Natural Language Processing held in NIT, Silchar, India. The workshop was organized virtually as a one-day event with two sessions. The first session included a keynote on "Computational Paninian Framework and Parsing Indian Languages" by Prof. Dipti Misra Sharma, followed by the technical sessions of the accepted papers. The afternoon session started with a keynote talk on "Parsing and its applications: Current Status and Future Perspectives" by Prof Marie-Catherine de Marneffe. The workshop was concluded with an insightful panel discussion on "Parsing and its applications: Current Status and Future Perspectives".

Annotated corpora are vital resources for deep learning application development and linguistic analyses. In the case of Indian languages, sufficient annotated quality data are not available publicly for researchers and developers to build upon. Although there are different levels of annotations, syntactically annotated corpora or treebank are very resourceful, especially for Indian languages, which are morphosyntactically rich. They are incorporated in downstream applications for various information extraction. Treebanks are being created for only a few Indian languages, and still, there is a high requirement for building more data in different domains while involving other languages. Indian languages do show language-specific complexities that require special attention. When a handful of languages representing different language families of India are ready with quality treebanks, those can be used to build resources for other languages using different approaches like transfer learning and multilingual learning. There is no way to start with or progress without annotated data; unsupervised approaches are not yet convincing enough, even for resourceful languages like English. Apart from all these justifications, we need to work together to make resources public and acceptable in a task like parsing and treebanking. Otherwise, we cannot create a meaningful and quality impact.

In this context, the PAIL-2021 workshop was organized with the following objects: (i) to bring researchers and developers together who work on treebanks, parsing, and related downstream natural language processing applications. (ii) to provide a platform for researchers to discuss Indian language-specific issues in the morphosyntactic analysis, and (iii) to encourage researchers to collaborate and create more annotated resources

Keynote speeches

Two keynote speeches were arranged to introduce two widely used frameworks to annotate the treebanks of Indian languages, namely, the Paninian framework and the Universal Dependencies framework. It was fortunate to get professors who could speak on these topics as keynote speakers.

Professor Emeritus Dipti Misra Sharma from the International Institute of Information Technology (IIIT), Gachibowli, Hyderabad, has spoken on "Computational Paninian Framework and Parsing Indian Languages". Her talk started with the advantages of Paninian Dependency frameworks in encoding syntactic and semantic relations between words in a sentence, especially for free word order languages like Indian languages. She also touched upon different approaches used for dependency parsing, starting from constraint-based parsing to the current neural network-based approaches. She also focused on treebanks developed in different Indian languages using available dependency frameworks. She briefly discussed the conversion of Paninian dependencies to Universal dependencies as well.

Professor Marie-Catherine de Marneffe from the Department of Linguistics, The Ohio State University, delivered the second talk on the topic entitled "Universal Dependencies: the good, the bad and the potential". She shared how the Universal Dependencies framework evolved and its philosophy first. Then she covered in detail the important concepts and the Universality of the framework and how it is used for various linguistics studies and the development of applications. Finally, she also set some directions for further improvements to the framework.

Technical sessions

Six papers were submitted for the PAIL-2021 workshop from India and Sri Lanka. Three experts reviewed each paper. Based on the outcomes, four papers were accepted for the workshop. While three of them covered the topic related to treebanking and parsing, one of them covered the application of treebanks.

The accepted papers covered a wide variety of topics. One paper was on Treebanking for extremely low-resource languages Braj and Magahi, and two papers were on syntactic parsing of Tamil and non-finite clauses in Telugu using feature-based Malt parser and rule-based approaches, respectively. In addition, there was a paper on Tamil grammar detection that touched on parsing usage.

Panel discussion

A panel discussion was also arranged to discuss a topic entitled “Parsing and its applications: Current Status and Future Perspectives". The primary object of this panel was to disseminate knowledge on parsing and treebanking and their importance and future. In addition, this was arranged to find out the challenges in creating relevant resources for Indian languages and the best practices. The following practitioners who participated in the panel discussion were well-experienced scholars in the area of parsing, tree-banking, and their applications:

- Professor Amba Kulkarni, Department for Sanskrit Studies, University of Hyderabad.
  Prof. Amba Kulkarni gave insights on parsing using Indian grammatical tradition and discussed the history of research in parsing Indian languages in general and Sanskrit in particular. She explained parsing complexities and the importance of including semantic information in parsing.
- Dr. Asif Ekbal, IIT Patna, India.
  Dr. Asif described the importance of parsing in various NLP applications and the optimization methods to integrate parsers in end-to-end neural network methods. He further discussed in detail how to develop multilingual treebanks with current techniques.
- Dr. Dan Zeman, Institute of Formal and Applied Linguistics, Charles University, Czech Republic.
  Dr. Dan Zemen, a parsing expert in Universal Dependencies (UD), discussed challenges in maintaining cross-lingual treebanks, mapping other treebanks with UD and building a new UD treebank. He explained the procedures of finding suitable texts with licence and the importance of documenting language-specific decisions.
- Dr. Ritesh Kumar, Dr Bhimrao Ambedkar University, India.
  Dr. Ritesh shared his experiences in treebank annotation for low-resourced languages such as Braj and Magahi. He also shared the language-specific issues that were encountered and how they are handled in the treebank building. He iterates that building treebank is a long-term activity that requires understanding the language-specific features, guidelines, and selecting the right annotation tool.

Acknowledgement: Organizers would like to thank everyone who supported us in organizing the first-ever workshop on parsing and its applications for the Indian Languages. Specifically, the ICON-2021 organizers and the technical programme committee members need to be acknowledged for the workshop facilitation and support in reviewing papers, respectively.

Kengatharaiyer Sarveswaran, University of Jaffna, Sri Lanka.
Parameswari Krishnamurthy, University of Hyderabad, India.
Pruthwik Mishra, IIIT-Hyderabad, India.

Organizers

- Kengatharaiyer Sarveswaran, Department of Computer Science, University of Jaffna, Sri Lanka.
- Parameswari Krishnamurthy, University of Hyderabad, India.
- Pruthwik Mishra, MT & NLP Lab, LTRC, IIIT-Hyderabad, India.

Technical Programme Committee

- Amba Kulkarni, University of Hyderabad, India
- Anand M Kumar, National Institute of Technology - Surathkal, India
- Ashwath Rao, MIT - Manipal, India
- Asif Ekbal, IIT Patna, India
- Braja Gopal Patra, Weill Cornell Medicine, USA
- Dhanalakshmi V, RV Government Arts college, India
- Asif Ekbal, IIT Patna, India
- Gihan Dias, University of Moratuwa, Sri Lanka
- Govindaru V, C-DIT, Thiruvanathapuram, India
- Irshad Ahmad Bhat, Active Intelligence LLP, India
- Malhar A Kulkarni, Indian Institute of Technology Bombay, India
- Muralikrishna SN, MIT - Manipal, India
- Ritesh Kumar, Dr Bhimrao Ambedkar University, India
- S Mahesan, University of Jaffna, Sri Lanka
- Rajendran Sankaravelayuthan, Amrita Vishwa Vidyapeetham, India
- Samar Hussain, IIT Delhi, India
- Sowmya Vajjala, National Research Council, Canada
- Surangika Ranathunga, University of Moratuwa, Sri Lanka
- Taraka Rama, University of North Texas, USA
- Uthayasanker Thayasivam, University of Moratuwa, Sri Lanka

Table of Contents

Developing Universal Dependencies Treebanks for Magahi and Braj
Mohit Raj, Shyam Ratan, Deepak Alok, Ritesh Kumar and Atul Kr. Ojha 1

Parsing Subordinate Clauses in Telugu using Rule-based Dependency Parser
P Sangeetha, Parameswari Krishnamurthy and Amba Kulkarni 12

Dependency Parsing in a Morphological rich language, Tamil
Vijay Sundar Ram and Sobha Lalitha Devi 20

Neural-based Tamil Grammar Error Detection
Dineskumar Murugesapillai, Anankan Ravinthirarasa, Gihan Dias and K Sarveswaran 27

Workshop Program

Thursday, December 16, 2021 - 09:30 to 16:50 (IST)

09:30–09:45 ***Morning sessions - Inaugural address***

09:45–10:45 *Keynote speech: Computational Paninian Framework and Parsing Indian Languages*
Professor Dipti Misra Sharma

10:45–11:15 *Developing Universal Dependencies Treebanks for Magahi and Braj*
Mohit Raj, Shyam Ratan, Deepak Alok, Ritesh Kumar and Atul Kr. Ojha

11:15–11:45 *Parsing Subordinate Clauses in Telugu using Rule-based Dependency Parser*
P Sangeetha, Parameswari Krishnamurthy and Amba Kulkarni

11:45–12:15 *Dependency Parsing in a Morphological rich language, Tamil*
Vijay Sundar Ram and Sobha Lalitha Devi

12:15–12:35 *Neural-based Tamil Grammar Error Detection*
Dineskumar Murugesapillai, Anankan Ravinthirarasa, Gihan Dias and Kengatharaiyer Sarveswaran

14:00–14:10 ***Evening sessions - Welcome address***

14:10–15:10 *Keynote speech: Universal Dependencies: the good, the bad and the potential*
Professor Marie-Catherine de Marneffe, Department of Linguistics, The Ohio State University.

15:10–16:45 *Panel discussion: Parsing and its applications: Current Status and Future Perspectives*
Prof. Amba Kulkarni, Dr. Asif Ekbal, Dr. Dan Zeman, and Dr. Ritesh Kumar