20th International Conference on Natural Language Processing (ICON 2023)
14 - 17 December 2023
Goa University

ICON 2023

BOOK OF ABSTRACTS

ORGANIZERS

Government of Goa
Directorate of Higher Education

Natural Language Processing
Association of India

GOA UNIVERSITY
गोय विद्यापीठ

SPONSORS

AUGNITO

HUBDATA

Adobe

Rian

tcs Research

ICON 2023 - Book of Abstracts

Table of Contents

Oral Presentations

9. **Active Learning Approach for Fine-Tuning Pre-Trained ASR Model for a Low-Resourced Language: A Case Study of Nepali**
   Rupak Raj Ghimire (Kathmandu University); Bal Krishna Bal (Department of Computer Science and Engineering, Kathmandu University, Nepal); Prakash Poudyal (Kathmandu University)
10. **Dispersed Hierarchical Attention Network for Machine Translation and Language Understanding on Long Documents with Linear Complexity**
    Ajay Mukund S (Anna University); K. S. Easwarakumar (Anna University)
11. **Analyzing Sentiment Polarity Reduction in News Presentation through Contextual Perturbation and Large Language Models**
    Alapan Kuila (IIT KHARAGPUR); Somnath Jena (IIT Kharagpur); Sudeshna Sarkar (IIT Kharagpur); Partha Chakrabarti (Indian Institute of Technology Kharagpur)
12. **NLI to the Rescue: Mapping Entailment Classes to Hallucination Categories in Abstractive Summarization**
    Naveen Badathala (Indian Institute of Technology Bombay); AshitaSaxena (Indian Institute of Technology Bombay); Pushpak Bhattacharyya (Indian Institute of Technology Bombay and Patna)
13. **Text Detoxification as Style Transfer in English and Hindi**
    Sourabrata Mukherjee (Charles University); Akanksha Bansal (Jawaharlal Nehru University); Atul Kr. Ojha (Data Science Institute, Unit for Linguistic Data, University of Galway); John P. McCrae (Insight Center for Data Analytics, National University of Ireland Galway); OndrejDusek (Charles University)
14. **Hindi Causal TimeBank: an Annotated Causal Event Corpus**
    Tanvi Kamble (International Institute of Information Technology, Hyderabad); Manish Shrivastava (International Institute of Information Technology, Hyderabad)
15. **Enriching Electronic Health Record with Semantic Features UtilisingPretrained Transformers**
    Lena Al Mutair (School of Computing, University of Leeds, Leeds, UK. Al-Imam Muhammad Ibn Saud Islamic University, Riyadh, SA); Eric Atwell (University of Leeds); Nishant Ravikumar (School of Computing, University of Leeds, Leeds, UK)
16. **Multilingual Multimodal Text Detection in Indo-Aryan Languages**
    Nihar Basisth (National Institute Of Technology, Silchar (NIT Silchar)); Eisha Halder (National Institute Of Technology, Silchar); Tushar Sachan (National Institute of Technology Silchar); Advaitha Vetagiri (National Institute of Technology Silchar); Partha Pakray (National Institute of Technology Silchar)
17. **Iterative Back Translation Revisited: An Experimental Investigation for Low-resource English Assamese Neural Machine Translation**
    Mazida Ahmed (Department of Information Technology, Gauhati University); Kishore Kashyap (Department of Information Technology , Gauhati University); Kuwali Talukdar (Gauhati University); Parvez Boruah (Gauhati University)
18. **Issues in the computational processing of Upamāalaṅkāra.**
    Bhakti Jadhav (Indian Institute of Technology Bombay); Amruta Barbadikar (University Of Hyderabad); Amba Kulkarni (University of Hyderabad); Malhar Kulkarni (IIT Bombay, India)

19. **Impacts of Approaches for Agglutinative-LRL Neural Machine Translation (NMT): A Case Study on Manipuri-English Pair**
    Gourashyam Moirangthem (Indian Institute of Information Technology (IIIT) Manipur); Lavinia Nongbri (Indian Institute of Information Technology (IIIT) Manipur); Samarendra Singh Salam (G.P. Women's College, Imphal); Kishorjit Nongmeikapam (Indian Institute of Information Technology(IIIT) Manipur)
20. **KITLM: Domain-Specific Knowledge InTegration into Language Models for Question Answering**
    Ankush Agarwal (IIT Bombay); Sakharam Gawade (Indian Institute of Technology Bombay); Amar Prakash Azad (IBM AI Research); Pushpak Bhattacharyya (Indian Institute of Technology Bombay and Patna)
21. **Neural Machine Translation for a Low Resource Language Pair: English-Bodo**
    Parvez Boruah (Gauhati University); Kuwali Talukdar (Gauhati University); Mazida Ahmed (Department of Information Technology, Gauhati University); Kishore Kashyap (Department of Information Technology , Gauhati University)
22. **Bi-Quantum Long Short-Term Memory for Part-of-Speech Tagging**
    Shyambabu Pandey (National Institute of Technology Silchar); Partha Pakray (National Institute of Technology Silchar)
23. **Sentiment Analysis for the Mizo Language: A Comparative Study of Classical Machine Learning and Transfer Learning Approaches**
    Mercy Lalthangmawii (National Institute of Technology Silchar); Thoudam Doren Singh (National Institute of Technology Silchar)
24. **Bidirectional Neural Machine Translation (NMT) using Monolingual Data for Khasi-English Pair**
    Lavinia Nongbri (IIIT Manipur); Gourashyam Moirangthem (Indian Institute of Information Technology (IIIT) Manipur); Samarendra Salam (G.P. Women's College, Imphal); Kishorjit Nongmeikapam (Indian Institute of Information Technology(IIIT) Manipur)
25. **Lost in Translation No More: Fine-tuned transformer-based models for CodeMix to English Machine Translation**
    Arindam Chatterjee (Wipro R&D, Lab45); Chhavi Sharma (Wipro Limited); Yashwanth V P (Wipro limited); Niraj Kumar (Wipro Limited); Ayush Raj (Wipro Limited); Asif Ekbal (IIT Patna)
26. **Automated System for Opinion Detection of Breathing Problem Discussions in Medical Forum Using Deep Neural Network**
    Somenath Nag Choudhury (IIT Patna); Asif Ekbal (Indian Institute of Technology, Patna)
27. **Effect of Pivot Language and Segment-Based Few-Shot Prompting for Cross-Domain Multi-Intent Identification in Low Resource Languages**
    Kathakali Mitra (BITS PILANI , Hyderabad Campus); Aditha Venkata Santosh Ashish (BITS PILANI , Hyderabad Campus); Soumya Teotia (BITS PILANI , Hyderabad Campus); Aruna Malapati (BITS PILANI , Hyderabad Campus)
28. **Towards Large Language Model driven Reference-less Translation Evaluation for English and Indian Language**
    Vandan Mujadia (IIIT-H); Pruthwik Mishra (IIIT, Hyderabad); Arafat Ahsan (IIIT Hyderabad); Dipti M. Sharma (IIITH)

29. **1-step Speech Understanding and Transcription Using CTC Loss**
    Karan Singla (SAIL, University of Southern California); Shahab Jalalvand (Interactions LLC); Yeon-Jun Kim (Interactions LLC); Andrej Ljolje (AT&T Labs - Research); Antonio Moreno Daniel (Interactions LLC); Srinivas Bangalore (Interactions Corp); Benjamin Stern (Interactions LLC)
30. **Consolidating Strategies for Countering Hate Speech Using Persuasive Dialogues**
    Sougata Saha (State University of New York at Buffalo); Rohini Srihari (University at Buffalo, SUNY)

Poster Presentations

1. **Konkani ASR**
   Swapnil Fadte (Computer Science and Technology, Goa University); Gaurish Thakkar (University of Zagreb); Jyoti Pawar (Goa University, Goa)
2. **Query-Based Summarization and Sentiment Analysis for Indian Financial Text by leveraging Dense Passage Retriever, RoBERTa, and FinBERT**
   Numair Shaikh (Department of Computer Engineering, SCTR's Pune Institute of Computer Technology); Jayesh Patil (Department of Computer Engineering, SCTR's Pune Institute of Computer Technology); Sheetal Sonawane (Department of Computer Engineering, SCTR's Pune Institute of Computer Technology)
3. **Bias Detection Using Textual Representation of Multimedia Contents**
   Karthik L Nagar (Accenture); Aditya Mohan Singh (Accenture Technology Labs); Sowmya Rasipuram (Accenture Technology Labs); Roshni Ramnani (Accenture Technology labs); Milind Savagaonkar (Accenture); Anutosh Maitra (Accenture)
4. **Rating YouTube Videos through Sentiment Analysis: A Case Study in Bangla-English code-mixed Situation**
   Arunava Kar (University of Hyderabad); Angshuman Jana (IIT Guwahati)
5. **Annotated and Normalized Causal Relation Extraction Corpus for Improving Health Informatics**
   Samridhi Dev (Jawaharlal Nehru University); Aditi Sharan (Jawaharlal Nehru University)
6. **T20NGD: Annotated corpus for news headlines classification in low resource language,Telugu.**
   CHINDUKURI MALLIKARJUNA (NIT-TIRUCHIRAPPALLI); Sangeetha Sivanesan (NIT-TIRUCHIRAPPALLI)
7. **Advancing Class Diagram Extraction from Requirement Text: A Transformer-Based Approach**
   Shweta . (Assistant Professor, Department of CSE, LNMIIT Jaipur); Suyash Mittal (Department of CSE, LNMIIT Jaipur); Suryansh Chauhan (Department of CSE, LNMIIT Jaipur)
8. **L3Cube-IndicNews: News-based Short Text and Long Document Classification Datasets in Indic languages**
   Aishwarya Mirashi (Pune Institute of Computer Technology); Srushti Sonavane (Indian); Purva Lingayat (Pune Institute of Computer Technology); Tejas Padhiyar (PICT); Raviraj Joshi (Indian Institute of Technology Madras)

9. **PoS to UPoS Conversion and Creation of UPoS Tagged Resources for Assamese Language**
   Kuwali Talukdar (Gauhati University); Prof. Shikhar Kumar Sarma (Gauhati University)
10. **Mitigating Abusive Comment Detection in Tamil Text: A Data Augmentation Approach with Transformer Model**
    Reshma Sheik (National Institute of Technology, Trichy); Raghavan Balanathan(National Institute of Technology, Trichy); S Jaya Nirmala (National Institute of Technology Tiruchirappalli)
11. **Dravidian Fake News Detection with Gradient Accumulation based Transformer Model**
    Eduri Raja (National Institute of Technology Silchar); Badal Soni (National Institute of Technology Silchar); Samir Kumar Borgohain (National Institute of Technology Silchar); Candy Lalrempuii (National Institute of Technology Silchar)
12. **Automatic Speech Recognition System for Malasar Language using Multilingual Transfer Learning**
    Basil K. Raju (Kerala University of Digital Sciences Innovation and Technology); Leena G. Pillai (Kerala University of Digital Sciences Innovation and Technology); Kavya Manohar (Kerala University of Digital Sciences Innovation and Technology); Elizabeth Sherly (Dr)
13. **Dy-poThon: A Bangla Sentence-Learning System for Children with Dyslexia**
    Dipshikha Podder (Indian Institute of Technology Kharagpur); Manjira Sinha (Assistant Professor, Indian Institute of Technology Kharagpur); Tirthankar Dasgupta (Tata Consultancy Services Ltd.); Anupam Basu (IIT Kharagpur)
14. **Mitigating Clickbait: An Approach to Spoiler Generation Using Multitask Learning**
    Sayantan Pal (State University of New York at Buffalo); Souvik Das (University at Buffalo); Rohini Srihari (University at Buffalo, SUNY)
15. **Comparing DAE-based and MASS-based UNMT: Robustness to Word-Order Divergence in English-->Indic Language Pairs**
    Tamali Banerjee (IIT Bombay); Rudra Murthy (IBM India Research Limited); Pushpak Bhattacharyya (Indian Institute of Technology Bombay and Patna)
16. **MahaSQuAD : Bridging Linguistic Divides in Marathi Question-Answering**
    Ruturaj Ghatage (Pune Institute of Computer Technology); Aditya Ashutosh Kulkarni (Pune Institute of Computer Technology); Rajlaxmi Patil (Pune Institute of Computer Technology); Sharvi Endait (Pune Institute of Computer Technology); Raviraj Joshi (Indian Institute of Technology Madras)
17. **CASM - Context and Something More in Lexical Simplification**
    Atharva Kumbhar (SCTR'S Pune Institute of Computer Technology); Sheetal Sonawane (SCTR'S Pune Institute of Computer Technology); Dipali Kadam (SCTR'S Pune Institute of Computer Technology); Prathamesh Mulay (Pune Institute Of Computer Technology)
18. **Improving the Evaluation of NLP Approaches for Scientific Text Annotation with Ontology Embedding-Based Semantic Similarity Metrics**
    Pratik Devkota (University of North Carolina at Greensboro); Somya D. Mohanty (United Health Care); Prashanti Manda (University of North Carolina at Greensboro)

19. **A Survey of using Large Language Models for Generating Infrastructure as Code**
    K Ganesh Srivatsa (International Institute of Information Technology, Hyderabad); Sabyasachi Mukhopadhyay (IIIT Hyderabad); Ganesh Katrapati (International Institute of Information Technology); Manish Shrivastava (International Institute of Information Technology Hyderabad)
20. **First Attempt at Building Parallel Corpora for Machine Translation of Northeast India's Very Low-Resource Languages**
    Atnafu Lambebo Tonja (InstitutoPolitécnico Nacional (IPN), Centro de Investigaciónen Computación (CIC)); Melkamu Mersha (University of Colorado ColoradoSprings); Ananya Kalita (University of Colorado); Olga Kolesnikova (Centro de Investigacionen Computacion del Instituto Politecnico Nacional); Jugal Kalita (University of Colorado)
21. **"Kurosawa": A Script Writer's Assistant**
    Prerak Gandhi (Indian Institute of Technology Bombay); Vishal Pramanik (IIT Bombay); Pushpak Bhattacharyya (Indian Institute of Technology Bombay and Patna)
22. **Text-2-Wiki: Summarization and Template-driven Article Generation**
    Jayant Panwar (International Institute of Information Technology); Radhika Mamidi (International Institute of Information Technology)
23. **Blind Leading the Blind: A Social-Media Analysis of the Tech Industry**
    Tanishq Chaudhary (International Institute of Information Technology Hyderabad); Pulak Malhotra (International Institute of Information Technology Hyderabad); Radhika Mamidi (International Institute of Information Technology Hyderabad); Ponnurangam Kumaraguru (International Institute of Information Technology Hyderabad)
24. **A Unified Multi task Learning Architecture for Hate Detection Leveraging User-based Information**
    Prashant Kapil (Indian Institute of Technology); Asif Ekbal (Indian Institute of Technology)
25. **Mytho-Annotator: An Annotation tool for Indian Hindu Mythology**
    Apurba Paul (Jadavpur University); Anupam Mondal (Institute of Engineering and Management); Sainik Mahata (Jadavpur University); Srijan Seal (JISCE); Prasun Sarkar (JISCE); Dipankar Das (Jadavpur University)
26. **Transformer-based Bengali Textual Emotion Recognition**
    Md. Atabuzzaman (Department of Computer Science, Virginia Tech); Maksuda Bilkis Baby (Hajee Mohammad Danesh Science and Technology University); Md Shajalal (Fraunhofer FIT)
27. **Citation-Based Summarization of Landmark Judgments**
    Purnima Bindal (University of Delhi); Vikas Kumar (University of Delhi); Vasudha Bhatnagar (University of Delhi); Parikshet Sirohi (University of Delhi); Ashwini Siwal (University of Delhi)
28. **Aspect and Opinion Term Extraction Using Graph Attention Network**
    Abir Chakraborty (Microsoft)
29. **Abstractive Hindi Text Summarization: A Challenge in a Low-Resource Setting**
    Daisy Lal (Lancaster University); Paul Rayson (Lancaster University); Krishna Singh (IIIT Allahabad); Uma Shanker Tiwary (Indian Institute of Information Technology Allahabad)

30. **Verb Categorisation for Hindi Word Problem Solving**
    Harshita Sharma (iiit.ac.in); Pruthwik Mishra (IIIT, Hyderabad); Dipti Sharma (IIIT, Hyderabad)
31. **ReviewCraft : A Word2Vec Driven System Enhancing User-Written Reviews**
    Gaurav Sawant (Goa University); Pradnya Bhagat (Goa University); Jyoti Pawar (Goa University,
32. **Intent Detection and Zero-shot Intent Classification for Chatbots**
    Sobha Lalitha Devi (AU-KBC Research Centre, Anna University); Pattabhi RK Rao (AU-KBC Research centre)
33. **Coreference Resolution Using AdapterFusion-based Multi-Task learning**
    Sobha Lalitha Devi (AU-KBC Research Centre, Anna University); Vijay Sundar Ram (AU-KBC, Anna University, Chennai); Pattabhi RK Rao (AU-KBC Research centre)
34. **Transfer learning in low-resourced MT: An empirical study**
    Sainik Mahata (Jadavpur University); Dipanjan Saha (Jadavpur University); Dipankar Das (Jadavpur University); Sivaji Bandyopadhyay (JADAVPUR UNIVERSITY, NIT SILCHAR)
35. **Transformer-based Nepali Text-to-Speech**
    Ishan Dongol (Kathmandu University); Bal Krishna Bal (Department of Computer Science and Engineering, Kathmandu University, Nepal)
36. **Infusing Knowledge into Large Language Models with Contextual Prompts**
    Kinshuk Vasisht (University of Delhi); Balaji Ganesan (IBM Research, India); Vikas Kumar (University of Delhi); Vasudha Bhatnagar (University of Delhi)
37. **Can Big Models Help Diverse Languages? Investigating Large Pretrained Multilingual Models for Machine Translation of Indian Languages**
    Telem Joyson Singh (IIT Guwahati); Sanasam Ranbir Singh (Indian Institute of Technology Guwahati); Priyankoo Sarmah (Indian Institute of Technology Guwahati)
38. **Revolutionizing Authentication: Harnessing Natural Language Understanding for Dynamic Password Generation and Verification**
    Akram Al-Rumaim (Goa University); Jyoti D. Pawar (Goa University)
39. **Leveraging Empathy, Distress, and Emotion for Accurate Personality Subtyping from Complex Human Textual Responses**
    Soumitra Ghosh (Fondazione Bruno Kessler (FBK), Italy); Tanisha Tiwari (Indian Institute of Technology Patna); Chetna Painkra (Indian institute of technology Patna); Gopendra Vikram Singh (IIT Patna); Asif Ekbal (Indian Institute of Technology Patna)
40. **A Baseline System for Khasi and Assamese Bidirectional NMT with Zero available Parallel Data : Dataset Creation and System Development**
    Kishore Kashyap (Department of Information Technology , Gauhati University); Kuwali Talukdar (Gauhati University); Mazida Ahmed (Department of Information Technology, Gauhati University); Parvez Boruah (Gauhati University)
41. **Parts of Speech (PoS) and Universal Parts of Speech (UPoS) Tagging: A Critical Review with Special Reference to Low Resource Languages**
    Kuwali Talukdar (Gauhati University); Prof. Shikhar Kumar Sarma (Gauhati University); Manash Pratim Bhuyan (Dibru College, Dibrugarh, Assam)
42. **Neural Machine Translation for Assamese-Bodo, a Low Resourced Indian Language Pair**
    Kuwali Talukdar (Gauhati University); Prof. Shikhar Kumar Sarma (Gauhati

University); Farha Naznin (Gauhati University); Kishore Kashyap (Department of Information Technology , Gauhati University); Mazida Ahmed (Department of Information Technology, Gauhati University); Parvez Boruah (Gauhati University)

43. **Attentive Fusion: A Transformer-based Approach to Multimodal Hate Speech Detection**
    Atanu Mandal (Jadavpur University); Gargi Roy (Optum Global Solutions Private Limited, Bengaluru, India); Amit Barman (Jadavpur University); Indranil Dutta (Jadavpur University); SudipNaskar (Jadavpur University)
44. **Handwritten Text Segmentation Using U-Net and Shuffled Frog-Leaping Algorithm with Scale Space Technique**
    MoumitaMoitra (NIT Durgapur); Sujan Kumar Saha (National Institute of Technology Durgapur)
45. **Identifying Correlation between Sentiment Analysis and Septic News Sentences Classification Tasks**
    Soma Das (Indian Institute of Information Technology Kalyani); Sagarika Ghosh (Indian Institute Of Information Technology Kalyani); Sanjay Chatterji (Indian Institute of Information Technology Kalyani)
46. **KT2: Kannada-Tulu Parallel Corpus Construction for Neural Machine Translation**
    Asha Hegde (Mangalore University); HosahalliLakshmaiahShashirekha (Mangalore University)
47. **Word Sense Disambiguation for Marathi language using Supervised Learning**
    RasikaRansing (DattaMeghe College of Engineering, Vidyalankar Institute of Technology); Archana Gulati (School of Business Management, NMIMS University, Mumbai)
48. **A Baseline System for Khasi and Assamese Bidirectional NMT with Zero available Parallel Data : Dataset Creation and System Development**
    Kishore Kashyap (Department of Information Technology , Gauhati University); Kuwali Talukdar (Gauhati University); Mazida Ahmed (Department of Information Technology, Gauhati University); Parvez Boruah (Gauhati University)
49. **Parts of Speech (PoS) and Universal Parts of Speech (UPoS) Tagging: A Critical Review with Special Reference to Low Resource Languages**
    KuwaliTalukdar (Gauhati University); Prof. Shikhar Kumar Sarma (Gauhati University); ManashPratimBhuyan (Dibru College, Dibrugarh, Assam)
50. **Neural Machine Translation for Assamese-Bodo, a Low Resourced Indian Language Pair**
    Kuwali Talukdar (Gauhati University); Prof. Shikhar Kumar Sarma (Gauhati University); Farha Naznin (Gauhati University); Kishore Kashyap (Department of Information Technology , Gauhati University); Mazida Ahmed (Department of Information Technology, Gauhati University); Parvez Boruah (Gauhati University)
51. **Attentive Fusion: A Transformer-based Approach to Multimodal Hate Speech Detection**
    Atanu Mandal (Jadavpur University); Gargi Roy (Optum Global Solutions Private Limited, Bengaluru, India); Amit Barman (Jadavpur University); Indranil Dutta (Jadavpur University); Sudip Naskar (Jadavpur University)

52. **Enhancing Telugu Part-of-Speech Tagging with Deep Sequential Models and Multilingual Embeddings**
    Sai Rishith Reddy Mangamuru (Amrita School of Artificial Intelligence, Coimbatore, Amrita Vishwa Vidyapeetham, India); Sai Prashanth Karnati (Amrita School of Artificial Intelligence, Coimbatore, Amrita Vishwa Vidyapeetham, India); Bala Karthikeya Sajja (Amrita School of Artificial Intelligence, Coimbatore, Amrita Vishwa Vidyapeetham, India); Divith Phogat (Amrita School of Artificial Intelligence, Coimbatore, Amrita Vishwa Vidyapeetham, India); Premjith B (Amrita School of Artificial Intelligence, Coimbatore, Amrita Vishwa Vidyapeetham, India)
53. **Unlocking Emotions in Text: A Fusion of Word Embeddings and Lexical Knowledge for Emotion Classification**
    Anjali Bhardwaj (South Asian University, New Delhi, India); Nesar Ahmad Wasi (South Asian University, New Delhi, India); Muhammad Abulaish (South Asian University)
54. **Convolutional Neural Networks can achieve binary bail judgement classification**
    Amit Barman (Jadavpur University); Devangan Roy (Jadavpur University); Debapriya Paul (Indian Institute of Engineering Science and Technology, Shibpur); Indranil Dutta (Jadavpur University); Shouvik Kumar Guha (Assistant Professor (Law), WBNUJS); Samir Karmakar (Jadavpur University); SudipNaskar (Jadavpur University)
55. **Multiset Dual Summarization for Incongruent News Article Detection**
    Sujit Kumar (Research Scholar, Indian Institute of Technology Guwahati); Rohan Jaiswal (Indian Institute of Technology Guwahati); Mohit Ram Sharma (Indian Institute of Technology Guwahati); SanasamRanbir Singh (Indian Institute of Technology Guwahati)
56. **Word Sense Disambiguation for Marathi language using Supervised Learning**
    RasikaRansing (DattaMeghe College of Engineering, Vidyalankar Institute of Technology); Archana Gulati (School of Business Management, NMIMS University, Mumbai)
57. **A comparative study of transformer and transfer learning MT models for English-Manipuri**
    Kshetrimayum Boynao Singh (National Institute of Technology Silchar); Ningthoujam Avichandra Singh (National Institute of Technology Silchar); Loitongbam Sanayai Meetei (National Institute of Technology Silchar); Ningthoujam Justwant Singh (National Institute Of Technology, Silchar); Thoudam Doren Singh (National Institute of Technology Silchar); Sivaji Bandyopadhyay (Jadavpur University, NIT Silchar)
58. **The Current Landscape of Multimodal Summarization**
    Atharva Kumbhar (SCTR'S Pune Institute of Computer Technology); Harsh Kulkarni (SCTR'S Pune Institute of Computer Technology); Atmaja Mali (SCTR'S Pune Institute of Computer Technology); Sheetal Sonawane (SCTR'S Pune Institute of Computer Technology); Prathamesh Mulay (SCTR'S Pune Institute of Computer Technology)
59. **Automated Answer Validation using Text Similarity**
    Balaji Ganesan (IBM Research); Arjun Ravikumar (INDIAN INSTITUTE OF SCIENCE); Lakshay Piplani (Independent); Rini Bhaumik ; Dhivya Padmanaban (Indian Institute of Science); Shwetha Narasimhamurthy (Independent Researcher); Chetan Adhikary (Tata Consultancy Services); Subhash Deshapogu (Independent Researcher)

60. **QeMMA: Quantum-Enhanced Multi-Modal Sentiment Analysis**
    Arpan Phukan (Indian Institute of Technology Patna); Asif Ekbal (Indian Institute of Technology Patna)
61. **Automatic Data Retrieval for Cross Lingual Summarization**
    Nikhilesh Bhatnagar (International Institute of Information Technology, Hyderabad); Ashok Urlana (TCS Research); Pruthwik Mishra (IIIT, Hyderabad); Vandan Mujadia (IIIT-H); Dipti Sharma (IIIT, Hyderabad)
62. **Cross-Lingual Fact Checking: Automated Extraction and Verification of Information from Wikipedia using References**
    Shivansh Subramanian (IIIT Hyderabad); Ankita Maity (IIIT Hyderabad); Aakash Jain (IIIT Hyderabad); Bhavyajeet Singh (IIIT Hyderabad); Harshit Gupta (International Institute of Information Technology, Hyderabad); Lakshya Khanna (IIIT Hyderabad); Vasudeva Varma (IIIT Hyderabad)
63. **Combining Pre trained Speech and Text Encoders for Continuous Spoken Language Processing**
    Karan Singla (SAIL, University of Southern California); Mahnoosh Mehrabani Interactions LLC); Daniel Pressel (member); Ryan Price (Interactions); Bhargav Srinivas Chinnari (Interactions LLC); Yeon-Jun Kim (Interactions LLC); Srinivas Bangalore (Interactions Corp)

Abstracts

Oral Presentations

IMAGINATOR: Pre-Trained Image+Text Joint Embeddings using Word-Level Grounding of Images

Varuna Krishna Kolla (University of Southern California); Suryavardan Suresh (New York University); Shreyash Mishra (IIITS); Sathyanarayanan Ramamoorthy (Carnegie Mellon Uiversity); Parth Patwa (University of California Los Angeles); Megha Chakraborty (University of South Carolina); Aman Chadha (Stanford University); Amitava Das (University of South Carolina); Amit Sheth (University of South Carolina)

varunakrishna.k19@iiits.in; suryavardan.s19@iiits.in; shreyash.m19@iiits.in; sathyanarayanan.r18@iiits.in; parthprasad.p17@iiits.in; meghac.tiya@gmail.com; aman@amanchadha.com; amitava.santu@gmail.com; amit@sc.edu

Abstract

Word embeddings, i.e., semantically meaningful vector representation of words, are largely influenced by the distributional hypothesis "You shall know a word by the company it keeps", whereas modern prediction-based neural network embeddings rely on design choices and hyperparameter optimization. Word embeddings like Word2Vec, GloVe etc. well capture the contextuality and real-world analogies but contemporary convolution-based image embeddings such as VGGNet, AlexNet, etc. do not capture contextual knowledge. The popular king-queen analogy does not hold true for most commonly used vision embeddings. In this paper, we introduce a pre-trained joint embedding (JE), named IMAGINATOR, trained on 21K distinct image objects. JE is a way to encode multimodal data into a vector space where the text modality serves as the grounding key, which the complementary modality (in this case, the image) is anchored with. IMAGINATOR encapsulates three individual representations: (i) object-object co-location, (ii) word-object co-location, and (iii) word-object correlation. These three ways capture complementary aspects of the two modalities which are further combined to obtain the final object-word JEs. Generated JEs are intrinsically evaluated to assess how well they capture the contextuality and real-world analogies. We also evaluate pre-trained IMAGINATOR JEs on three downstream tasks: (i) image captioning, (ii) Image2Tweet, and (iii) text-based image retrieval. IMAGINATOR establishes a new standard on the aforementioned downstream tasks by outperforming the current SoTA on all the selected tasks. The code is available at https://github.com/varunakk/IMAGINATOR.

Evaluating user preferences in Hindi Text-to-Speech

Bharat Gupta (Ministry of Electronics and Information Technology Government of India)

bharatg@gov.in

Abstract

Hindi holds the distinction of being the fourth most extensively spoken first language globally. It serves as an official language in India, encompassing several states within the country. Hindi also has a number of dialects that are scattered over the entire Hindi-speaking region. There hasn't been a phonological comparison of Hindi dialects. Text-To-Speech (TTS) systems can only be evaluated effectively using the mean opinion score (MOS) and degradation mean opinion score (DMOS) as recommended metrics for synthesized speech quality. These subjective metrics are the most widely used to assess speech synthesis. During the evaluation phase, numerous assessors from different locations tend to exhibit a bias towards

their prosodic style. They often feel more comfortable when both speaking and listening in their native language with a prosodic manner. In this report, we studied the Hindi region's evaluators' preferences while evaluating the Hindi TTS system due to language's dialects and prosody, the study focuses on the influence of language dialects and prosody on the Degradation Mean Opinion Score (DMOS) and overall system performance. The current research is to discover the preferences and appropriate weightage of the dialects while evaluating the performance of Hindi TTS. Through a comparative analysis and an exploration of the details and recommended weights assigned to different dialects based on preferences, this research examines the variations in scores offered by different evaluators. The current research investigates various patterns of dialects in terms of character and word variations. The words variations have been organized by building Chhattisgarhi's& Haryanvi's word dictionary w.r.t Hindi. The evaluation score has also been analyzed by conducting evaluation testing on Hindi TTS having characters/words variations of the dialects. The W3C SSML (Speech Synthesis Markup Language) has been built for the implementation of various patterns of dialects on the TTS system. The current approach has been made for the development of dialect based TTS that is not currently available in the system.

Multi-Hop Relation Aware Representations for Inductive Knowledge Graphs

AniruddhaBala (Samsung); Ankit Sharma (Samsung Electronics); Shlok Sharma (Samsung); PinakiBhaskar (Advanced Technology Lab (ATL), Samsung R&D Institute India - Bangalore (SRI-B))

aniruddha127@gmail.com; mr.ankit019@gmail.com; shlok.sharma@samsung.com; pinaki.b@samsung.com

Abstract

Recent knowledge graph (KG) embedding methods explore parameter-efficient representations for large-scale KGs. These techniques learn entity representation using a fixed size vocabulary. Such a vocabulary consists of all the relations and a small subset of the full entity set, referred to as anchors. An entity is hence expressed as a function of reachable anchors and immediate relations. The performance of these methods is, therefore, largely dependent on the entity tokenization strategy. Especially in inductive settings, the representation capacity of these embeddings is limited due to the absence of anchor entities, as unseen entities have no connection with training graph entities. \\ In this work, we propose a novel entity tokenization strategy that tokenizes an entity into a set of anchors based on relation similarity and relational paths. Our model MH-RARe overcomes the challenge of unseen entities not being directly connected to the anchors by selecting informative anchors from the training graph using relation similarity. Experiment results show that our model outperforms the baselines on multiple datasets for inductive knowledge graph completion task, attaining upto 5 % improvement, while maintaining parameter efficiency.

Pronunciation-Aware Syllable Tokenizer for Nepali Automatic Speech Recognition System

Rupak Raj Ghimire (Kathmandu University); Bal Krishna Bal (Department of Computer Science and Engineering, Kathmandu University, Nepal); BalaramPrasain (Central Department of Linguistics, Tribhuvan University, Nepal); Prakash Poudyal (Kathmandu University)

rughimire@gmail.com; bal@ku.edu.np; balaram.prasain@cdl.tu.edu.np; prakash@ku.edu.np

Abstract

The Automatic Speech Recognition (ASR) has come up with significant advancements over the course of several decades, transitioning from a rule-based method to a statistical approach, and ultimately to the use of end-to-end (E2E) frameworks. This phenomenon continues with the progression of machine learning and deep learning methodologies. The E2E approach for ASR has demonstrated predominant success in the case of resourceful languages with larger annotated corpus. However, the accuracy is quite low for low-resourced languages such as Nepali. In this regard, language-specific tools such as tokenizers seem to

play a vital role in improving the performance of the E2E model for low-resourced languages like Nepali. In this paper, we propose a pronunciation-aware syllable tokenizer for the Nepali language which improves the results of the E2E model. Our experiment confirm that the introduction of the proposed tokenizer yields better performance with the Character Error Rate (CER) 8.09% compared to other language-independent tokenizers.

Neural language model embeddings for Named Entity Recognition: A study from language perspective

MuskaanMaurya (The English and Foreign Languages University); Anupam Mandal (Centre for AI & Robotics); Manoj Maurya (Centre for AI & Robotics); Naval Gupta (Centre for AI & Robotics); SomyaNayak (The English & Foreign Languages University)
muskaan.maurya06@gmail.com; amandal.cair@gov.in; manoj.cair@gov.in; naval.gupta.cair@gov.in; somyanayak@efluniversity.ac.in

Abstract
Named entity recognition (NER) models based on neural language models (LMs) exhibit state-of-the-art performance. However, such LMs have not been studied in detail with respect to finer language-related aspects in the context of NER tasks. Such a study is important for the effective application of these models for cross-lingual and multilingual NER tasks. In this study, we examine the effects of script, vocabulary sharing, content, and pooling of multilanguage training data for building NER models. It is observed that monolingual BERT embeddings show the highest recognition accuracy among all transformer-based LMs for monolingual NER models. It is also seen that vocabulary sharing and data augmentation with foreign NEs are most effective for cross-lingual NER models. Multilingual NER models trained by pooling data from similar languages can address training data inadequacy and exhibit performance close to that of monolingual models trained with adequate NER-tagged data of a single language.

Understanding behaviour of large language models for short-term and long-term fairness scenarios

TalhaChafekar (KJ Somaiya College of Engineering); Aafiya Hussain (KJ Somaiya College of Engineering); Chon In Cheong (University of Cambridge)
talha1503@gmail.com; aafiya.h@somaiya.edu; cic34@cam.ac.uk

Abstract
Large language models (LLMs) have become increasingly accessible online, thus they can be easily used to generate synthetic data for technology. With the rising capabilities of LLMs, their applications span across many domains. With its increasing use for automating tasks, it is crucial to understand the fairness notions harboured by these models. Our work aims to explore the consistency and behaviour of GPT-3.5, GPT-4 in both short-term and long-term scenarios through the lens of fairness. Additionally, the search for an optimal prompt template design for equalized opportunities has been investigated in this study. In the short-term scenario for the German Credit dataset, an intervention to a key feature recorded an increase in loan rejection rate by 37.15% for GPT-3.5 and 49.52% for GPT-4. In the long-term scenario for ML fairness gym, adding extra information about the environment to the prompts has shown no improvement to the prompt with minimal information in terms of final credit distributions. However, adding extra features to the prompt has increased the profit rate by 6.41% (from 17.2% to 23.6%) compared to a baseline maximum-reward classifier with compromising group-level recall rates.

Identifying Intent-Sentiment Co-reference from Legal Utterances

PinakiKarkun (Jadavpur University); Dipankar Das (Jadavpur University)
Pinaki.Karkun@gmail.com; dipankar.dipnil2005@gmail.com

Abstract
Co-reference is always treated as one of challenging tasks under natural language processing and has been explored only in the domain of anaphora resolution to an extent. However, the benefit of it to identify the relations between multiple entities in a single context can be explored better while we aim to identify intent and sentiment from the utterances of a dialogue or conversation. The utilization of co-reference becomes more elegant while tracking users' intents with respect to their corresponding sentiments explored in a specialized domain like judiciary. Thus, in the present attempt, we have identified not only intent and sentiment expressions at token level in an individual manner, we also classified the utterances and identified the co-reference between intent and sentiment entities in utterance level context. Last but not the least, the deep learning algorithms have shown improvements over traditional machine learning in all cases.

An Annotated Corpus for Realis Event Detection in Short Stories Written in English and Low Resource Assamese Language

Chaitanya Kirti (Indian Institute of Technology Guwahati); Pankaj Choudhury (Indian Institute of Technology Guwahati); Ashish Anand (Indian Institute of Technology Guwahati); PrithwijitGuha (Indian Institute of Technology Guwahati)
ckirti@iitg.ac.in; pankajchoudhury@iitg.ac.in; anand.ashish@iitg.ac.in; pguha@iitg.ac.in

Abstract
This paper presents an annotated corpora of Assamese and English short stories for event trigger detection. This marks a pioneering endeavor in short stories, contributing to developing resources for this genre, especially in the low-resource Assamese language. In the process, 200 short stories were manually annotated in both Assamese and English. The dataset was evaluated and several models were compared for predicting events that are actually happening, i.e., realis events. However, it is expensive to develop manually annotated language resources, especially when the text requires specialist knowledge to interpret. In this regard, TagIT, an automated event annotation tool, is introduced. TagIT is designed to facilitate our objective of expanding the dataset from 200 to 1,000. The best-performing model was employed in TagIT to automate the event annotation process. Extensive experiments were conducted to evaluate the quality of the expanded dataset. This study further illustrates how the combination of an automatic annotation tool and human-in-the-loop participation significantly reduces the time needed to generate a high-quality dataset.

Active Learning Approach for Fine-Tuning Pre-Trained ASR Model for a Low-Resourced Language: A Case Study of Nepali

Rupak Raj Ghimire (Kathmandu University); Bal Krishna Bal (Department of Computer Science and Engineering, Kathmandu University, Nepal); Prakash Poudyal (Kathmandu University)
rughimire@gmail.com; bal@ku.edu.np; prakash@ku.edu.np

Abstract
Fine tuning of the pre-trained language model is a technique which can be used to enhance the technologies of low-resourced languages. The unsupervised approach can fine-tune any pre-trained model with minimum or even no language-specific re- sources. It is highly advantageous, particularly for languages that possess limited computational resources. We present a novel approach for fine-tuning a pre-trained Automatic Speech Recognition (ASR) model that is suitable for low resource languages. Our

methods involves iterative fine-tuning of pre-trained ASR model. mms-1b is selected as the pretrained seed model for fine-tuning. We take the Nepali language as a case study for this research work. Our approach achieved a CER of 6.77%, outperforming all previously recorded CER values for the Nepali ASR Systems.

Dispersed Hierarchical Attention Network for Machine Translation and Language Understanding on Long Documents with Linear Complexity

Ajay Mukund S (Anna University); K. S. Easwarakumar (Anna University)
ajaymukund1998@gmail.com; easwara@annauniv.edu

Abstract
Transformers, being the forefront of Natural Language Processing and a pioneer in the recent developments, we tweak the very fundamentals of the giant Deep Learning model in this paper. For long documents, the conventional Full Self-Attention exceeds the compute power and the memory requirement as it scales quadratically. Instead, if we use a Local Self-Attention using a sliding window, we lose the global context present in the input document which can impact the performance of the task in hand. For long documents (ranging from 500 to 16K tokens), the proposed Dispersed Hierarchical Attention component captures the local context using a sliding window and the global context using a linearly-scaled dispersion approach. This achieves O(N) linear complexity, where N is the length of the input sequence or document.

Analyzing Sentiment Polarity Reduction in News Presentation through Contextual Perturbation and Large Language Models

AlapanKuila (IIT KHARAGPUR); Somnath Jena (IIT Kharagpur); Sudeshna Sarkar (IIT Kharagpur); ParthaChakrabarti (Indian Institute of Technology Kharagpur)
alapan.cse@gmail.com; somnathjena.2011@gmail.com; sudeshna@cse.iitkgp.ac.in; ppchak@cse.iitkgp.ac.in

Abstract
In today's media landscape, where news outlets play a pivotal role in shaping public opinion, it is imperative to address the issue of sentiment manipulation within news text. News writers often inject their own biases and emotional language, which can distort the objectivity of reporting. This paper introduces a novel approach to tackle this problem by reducing the polarity of latent sentiments in news content. Drawing inspiration from adversarial attack-based sentence perturbation techniques and a prompt-based method using ChatGPT, we employ transformation constraints to modify sentences while preserving their core semantics. Using three perturbation methods—replacement, insertion, and deletion—coupled with a context-aware masked language model, we aim to maximize the desired sentiment score for targeted news aspects through a beam search algorithm. Our experiments and human evaluations demonstrate the effectiveness of these two models in achieving reduced sentiment polarity with minimal modifications while maintaining textual similarity, fluency, and grammatical correctness. Comparative analysis confirms the competitive performance of the adversarial attack-based and prompt-based methods, offering a promising solution to foster more objective news reporting and combat emotional language bias in the media.

NLI to the Rescue: Mapping Entailment Classes to Hallucination Categories in Abstractive Summarization

Naveen Badathala (Indian Institute of Technology Bombay); AshitaSaxena (Indian Institute of Technology Bombay); Pushpak Bhattacharyya (Indian Institute of Technology Bombay and Patna)
naveenbadathala@cse.iitb.ac.in; ashitasaxena@cse.iitb.ac.in; pushpakbh@gmail.com

Abstract
In this paper, we detect hallucinations in summaries generated by abstractive summarization models. We focus on three types of hallucination viz. intrinsic, extrinsic, and non-hallucinated. The method used for detecting hallucination is based on textual entailment. Given a premise and a hypothesis, textual entailment classifies the hypothesis as contradiction, neutral, or entailment. These three classes of textual entailment are mapped to intrinsic, extrinsic, and non-hallucinated respectively. We fine-tune a RoBERTa-large model on NLI datasets and use it to detect hallucinations on the XSumFaith dataset. We demonstrate that our simple approach using textual entailment outperforms the existing factuality inconsistency detection systems by 12% and we provide insightful analysis of all types of hallucination. To advance research in this area, we create and release a dataset, XSumFaith++, which contains balanced instances of both hallucinated and non-hallucinated summaries.

Text Detoxification as Style Transfer in English and Hindi

Sourabrata Mukherjee (Charles University); Akanksha Bansal (Jawaharlal Nehru University); Atul Kr. Ojha (Data Science Institute, Unit for Linguistic Data, University of Galway); John P. McCrae (Insight Center for Data Analytics, National University of Ireland Galway); OndrejDusek (Charles University)
soura1990@gmail.com; akanksha.bansal15@gmail.com; shashwatup9k@gmail.com; john@mccr.ae; odusek@ufal.mff.cuni.cz

Abstract
This paper focuses on text detoxification, i.e., automatically converting toxic text into non-toxic text. This task contributes to safer and respectful online communication and can be considered as a Text Style Transfer (TST) task, where the text's style changes while its content is preserved. We present three approaches: (i) knowledge transfer from a similar task (ii) multi-task learning approach, combining sequence-to-sequence modeling with various toxicity classification tasks, and (iii) delete and reconstruct approach. To support our research, we utilize a dataset provided by Dementieva et al. (2021), which contains multiple versions of detoxified texts corresponding to toxic texts. In our experiments, we selected the best variants through expert human annotators, creating a dataset where each toxic sentence is paired with a single, appropriate detoxified version. Additionally, we introduced a small Hindi parallel dataset, aligning with a part of the English dataset, suitable for evaluation purposes. Our results demonstrate that our approach effectively balances text detoxification while preserving the actual content and maintaining fluency.

Hindi Causal TimeBank: an Annotated Causal Event Corpus

TanviKamble (International Institute of Information Technology, Hyderabad); Manish Shrivastava (International Institute of Information Technology, Hyderabad)
tanvi.kamble@research.iiit.ac.in; m.shrivastava@iiit.ac.in

Abstract
Events and states have gained importance in NLP and information retrieval for being semantically rich temporal and spatial information indicators. Event causality helps us identify which events are necessary

for another event to occur. The cause-effect event pairs can be relevant for multiple NLP tasks like question answering, summarization, etc. Multiple efforts have been made to identify causal events in documents but very little work has been done in this field in the Hindi language. We create an annotated corpus for detecting and classifying causal event relations on top of the Hindi Timebank, the `Hindi Causal Timebank' (Hindi CTB). We introduce semantic causal relations like Purpose, Reason, and Enablement inspired from Bejan's annotation scheme and add some special cases particular to Hindi language.

Enriching Electronic Health Record with Semantic Features Utilising Pretrained Transformers

Lena AlMutair (School of Computing, University of Leeds, Leeds, UK. Al-Imam Muhammad Ibn Saud Islamic University,Riyadh, SA); Eric Atwell (University of Leeds); NishantRavikumar (School of Computing, University of Leeds, Leeds, UK)
lena.almutair@gmail.com; e.s.atwell@leeds.ac.uk; n.ravikumar@leeds.ac.uk

Abstract
Electronic Health Records (EHRs) have revolutionised healthcare by enhancing patient care and facilitating provider communication. Nevertheless, the efficient extraction of valuable information from EHRs poses challenges, primarily due to the overwhelming volume of unstructured data, the wide variability in data formats, and the lack of standardised labels. Leveraging deep learning and concept embeddings, we address the gap in context-aware systems for EHRs. The proposed solution was evaluated on the MIMIC III dataset and demonstrated superior performance compared to other methodologies. We addressed the positive impact of the latent feature combined with the note representation in four different settings. Model performance was evaluated using a case study conducted with BertScore, assessing precision, recall, and F1 scores. The model excels in Medical Natural Language Inference (MedNLI) with an 89.3% accuracy, further boosted to 90.5% through retraining the embeddings using International Classification of Diseases (ICD) codes, which we formally designate as ClinicNarrIR. The ClinicNarrIR was tested with 1000 randomly sampled notes, achieving an NDCG@10 score of approximately 0.54 with accuracy@10 of 0.85. The study also demonstrates a high correlation between the results produced by the proposed representation and medical coders. Notably, in all evaluation cases, the optimal base pretrainedmodel that emerged was BlueBERT.

Multilingual Multimodal Text Detection in Indo-Aryan Languages

NiharBasisth (National Institute Of Technology, Silchar (NIT Silchar)); Eisha Halder (National Institute Of Technology, Silchar); TusharSachan (National Institute of Technology Silchar); AdvaithaVetagiri (National Institute of Technology Silchar); ParthaPakray (National Institute of Technology Silchar)
niharjyotibasisth@gmail.com; eishashalder@gmail.com; tusharsachan2014@gmail.com; advaithavetagiri@gmail.com; parthapakray@gmail.com

Abstract
Multi-language text detection and recognition in complex visual scenes is an essential yet challenging task. Traditional pipelines relying on optical character recognition (OCR) often fail to generalize across different languages, fonts, orientations and imaging conditions. This work proposes a novel approach using the YOLOv5 object detection model architecture for multilanguage text detection in images and videos. We curate and annotate a new dataset of over 4,000 scene text images across 4 Indian languages and use specialized data augmentation techniques to improve model robustness. Transfer learning from a base YOLOv5 model pretrained on COCO is combined with tailored optimization strategies for multi-language text detection. Our approach achieves state-of-theart performance, with over 90% accuracy on

multi-language text detection across all four languages in our test set. We demonstrate the effectiveness of fine-tuning YOLOv5 for generalized multi-language text extraction across diverse fonts, scales, orientations, and visual contexts. Our approach's high accuracy and generalizability could enable numerous applications involving multilingual text processing from imagery and video.

Iterative Back Translation Revisited: An Experimental Investigation for Low-resource English Assamese Neural Machine Translation

Mazida Ahmed (Department of Information Technology, Gauhati University); Kishore Kashyap (Department of Information Technology , Gauhati University); KuwaliTalukdar (Gauhati University); Parvez Boruah (Gauhati University)
14mazida.ahmed@gmail.com; kb.guwahati@gmail.com; kuwalitalukdar@gmail.com; parvezaziz70@gmail.com

Abstract
Back Translation has been an effective strategy to leverage monolingual data both on the source and target sides. Research have opened up several ways to improvise the procedure, one among them is iterative back translation where the monolingual data is repeatedly translated and used for re-training for the model enhancement. Despite its success, iterative back translation remains relatively unexplored in low-resource scenarios, particularly for rich Indic languages. This paper presents a comprehensive investigation into the application of iterative back translation to the low-resource English-Assamese language pair. A simplified version of iterative back translation is presented. This study explores various critical aspects associated with back translation, including the balance between original and synthetic data and the refinement of the target (backward) model through cleaner data retraining. The experimental results demonstrate significant improvements in translation quality. Specifically, the simplistic approach to iterative back translation yields a noteworthy +6.38 BLEU score improvement for the English-Assamese translation direction and a +4.38 BLEU score improvement for the Assamese-English translation direction. Further enhancements are further noticed when incorporating higher-quality, cleaner data for model retraining highlighting the potential of iterative back translation as a valuable tool for enhancing low-resource neural machine translation (NMT).

Issues in the computational processing of Upamāalaṅkāra.

Bhakti Jadhav (Indian Institute of Technology Bombay); AmrutaBarbadikar (University Of Hyderabad); Amba Kulkarni (University of Hyderabad); malharkulkarni (IIT Bombay, India)
bhaktij96@gmail.com; amruta.barbadikar@gmail.com; ambapradeep@gmail.com; malharku@gmail.com

Abstract
Processing and understanding of figurative speech is a challenging task for computers as well as humans. In this paper, we present a case of Upamāalaṅkāra (simile). The verbal cognition of the Upamāalaṅkāra by a human is presented as a dependency tree, which involves the identification of various components such as upamāna (vehicle), upameya (topic), sādhāraṇadharma (common property) and upamādyotaka (word indicating similitude).This involves the repetition of elliptical elements. Further, we show, how the same dependency tree may be represented without any loss of information, even without repetition of elliptical elements. Such a representation would be useful for the computational processing of the alaṅkāras.

Impacts of Approaches for Agglutinative-LRL Neural Machine Translation (NMT): A Case Study on Manipuri-English Pair

Gourashyam Moirangthem (Indian Institute of Information Technology (IIIT) Manipur); Lavinia Nongbri (Indian Institute of Information Technology (IIIT) Manipur); Samarendra Singh Salam (G.P. Women's College, Imphal); Kishorjit Nongmeikapam (Indian Institute of Information Technology(IIIT) Manipur)
mgourashyam@gmail.com; laviniangbri@gmail.com; samar.crypt@gmail.com; kishorjit@iiitmanipur.ac.in

Abstract
Neural Machine Translation (NMT) is known to be extremely challenging for Low-Resource Languages (LRL) with complex morphology. This work deals with the NMT of a specific LRL called Manipuri/Meeteilon, which is a highly agglutinative language where words have extensive suffixation with limited prefixation. The work studies and discusses the impacts of approaches to mitigate the issues of NMT involving agglutinative LRL in a strictly low-resource setting. The research work experimented with several methods and techniques including subword tokenization, tuning of the self-attention-based NMT model, utilization of monolingual corpus by iterative back-translation, embedding-based sentence filtering for back translation. This research work in the strictly low resource setting of only 21204 training sentences showed remarkable results with a BLEU score of 28.17 for Manipuri to English translation.

KITLM: Domain-Specific Knowledge InTegration into Language Models for Question Answering

Ankush Agarwal (IIT Bombay); Sakharam Gawade (Indian Institute of Technology Bombay); Amar Prakash Azad (IBM AI Research); Pushpak Bhattacharyya (Indian Institute of Technology Bombay and Patna)
ankush98.12@gmail.com; sakharamg@cse.iitb.ac.in; amarazad@gmail.com; pushpakbh@gmail.com

Abstract
Large language models (LLMs) have demonstrated remarkable performance in a wide range of natural language tasks. However, as these models continue to grow in size, they face significant challenges in terms of computational costs. Additionally, LLMs often lack efficient domain-specific understanding, which is particularly crucial in specialized fields such as aviation and healthcare. To boost the domain-specific understanding, we propose KITLM, a novel knowledge base integration approach into language model through relevant information infusion. By integrating pertinent knowledge, not only the performance of the language model is greatly enhanced, but the model size requirement is also significantly reduced while achieving comparable performance. Our proposed knowledge-infused model surpasses the performance of both GPT-3.5-turbo and the state-of-the-art knowledge infusion method, SKILL, achieving over 1.5 times improvement in exact match scores on the MetaQA. KITLM showed a similar performance boost in the aviation domain with AeroQA. The drastic performance improvement of KITLM over the existing methods can be attributed to the infusion of relevant knowledge while mitigating noise. In addition, we release two curated datasets to accelerate knowledge infusion research in specialized fields: a) AeroQA, a new benchmark dataset designed for multi-hop question-answering within the aviation domain, and b) Aviation Corpus, a dataset constructed from unstructured text extracted from the National Transportation Safety Board reports. Our research contributes to advancing the field of domain-specific language understanding and showcases the potential of knowledge infusion techniques in improving the performance of language models on question-answering.

Neural Machine Translation for a Low Resource Language Pair: English-Bodo

Parvez Boruah (Gauhati University); Kuwali Talukdar (Gauhati University); Mazida Ahmed (Department of Information Technology, Gauhati University); Kishore Kashyap (Department of Information Technology , Gauhati University)
parvezaziz70@gmail.com; kuwalitalukdar@gmail.com; 14mazida.ahmed@gmail.com; kb.guwahati@gmail.com

Abstract
This paper represents a work done on Neural Machine Translation for English and Bodo language pair. Bodo is a language mostly spoken in North Eastern area of India. This work of machine translation is done on a relatively small size of parallel data as there is less parallel corpus available for English-Bodo pair. Corpus is generally taken from available source National Platform of Language Technology(NPLT), Data Management Unit(DMU), Mission Bhashini, Ministry of Electronics and Information Technology and also generated internally in-house. Tokenization of raw text is done using IndicNLP library and mosesdecoder for Bodo and English respectively. Subword tokenization is performed by using BPE(Byte Pair Encoder) , Sentence piece and Wordpiecesubword. Experiments have been done on two different vocab size of 8000 and 16000 on a total of around 92410 parallel sentences. Two standard transformer encoder and decoder models with varying number of layers and hidden size are built for training the data using OpenNMT-py framework. The results are recorded based on the BLEU score on an additional testset for evaluating the performance. The highest BLEU score of 11.01 and 14.62 are achieved on the testset for English to Bodo and Bodo to English translation respectively.

Bi-Quantum Long Short-Term Memory for Part-of-Speech Tagging

Shyambabu Pandey (National Institute of Technology Silchar); ParthaPakray (National Institute of Technology Silchar)
shyambabu21_rs@cse.nits.ac.in; parthapakray@gmail.com

Abstract
Natural language processing (NLP) is a subfield of artificial intelligence that enables computer systems to understand and generate human language. NLP tasks involved machine learning and deep learning methods for processing the data. Traditional applications utilize massive datasets and resources to perform NLP applications, which is challenging for classical systems. On the other hand, Quantum computing has emerged as a promising technology with the potential to address certain computational problems more efficiently than classical computing in specific domains. In recent years, researchers have started exploring the application of quantum computing techniques to NLP tasks. In this paper, we propose a quantum-based deep learning model, Bi-Quantum long short-term memory (BiQLSTM). We apply POS tagging using the proposed model on social media code-mixed datasets.

Sentiment Analysis for the Mizo Language: A Comparative Study of Classical Machine Learning and Transfer Learning Approaches

Mercy Lalthangmawii (National Institute of Technology Silchar); Thoudam Doren Singh (National Institute of Technology Silchar)
mercy_pg_22@cse.nits.ac.in; thoudam.doren@gmail.com

Abstract
Sentiment analysis, a subfield of natural language processing (NLP) has witnessed significant advancements in the analysis of user-generated contents across diverse languages. However, its application to low-resource languages remains a challenge. This research addresses this gap by conducting a comprehensive sentiment analysis experiment in the context of the Mizo language, a low-

resource language predominantly spoken in the Indian state of Mizoram and neighboring regions. Our study encompasses the evaluation of various machine learning models including Support Vector Machine (SVM), Decision Tree, Random Forest, K-Nearest Neighbor (K-NN), Logistic Regression and transfer learning using XLM-RoBERTa. The findings reveal the suitability of SVM as a robust performer in Mizo sentiment analysis demonstrating the highest F-1 Score and Accuracy among the models tested. XLM-RoBERTa, a transfer learning model exhibits competitive performance highlighting the potential of leveraging pre-trained multilingual models in low-resource language sentiment analysis tasks. This research advances our understanding of sentiment analysis in low-resource languages and serves as a stepping stone for future investigations in this domain.

Bidirectional Neural Machine Translation (NMT) using Monolingual Data for Khasi-English Pair

Lavinia Nongbri (IIIT Manipur); Gourashyam Moirangthem (Indian Institute of Information Technology (IIIT) Manipur); Samarendra Salam (G.P. Women's College, Imphal); Kishorjit Nongmeikapam (Indian Institute of Information Technology(IIIT) Manipur)
laviniangbri@gmail.com; mgourashyam@gmail.com; samar.crypt@gmail.com; kishorjit@iiitmanipur.ac.in

Abstract
Due to a lack of parallel data, low-resource language machine translation has been unable to make the most of Neural Machine Translation. This paper investigates several approaches as to how low-resource Neural Machine Translation can be improved in a strictly low-resource setting, especially for bidirectional Khasi-English language pairs. The back-translation method is used to expand the parallel corpus using monolingual data. The work also experimented with subword tokenizers to improve the translation accuracy for new and rare words. Transformer, a cutting-edge NMT model, serves as the backbone of the bidirectional Khasi-English machine translation. The final Khasi-to-English and English-to-Khasi NMT models trained using both authentic and synthetic parallel corpora show an increase of 2.34 and 3.1 BLEU scores, respectively, when compared to the models trained using only authentic parallel dataset.

Lost in Translation No More: Fine-tuned transformer-based models for CodeMix to English Machine Translation

Arindam Chatterjee (Wipro R&D, Lab45); Chhavi Sharma (Wipro Limited); Yashwanth V P (Wipro limited); Niraj Kumar (Wipro Limited); Ayush Raj (Wipro Limited); Asif Ekbal (IIT Patna)
arindam.chatterjee4@wipro.com; chhavi.sharma5@wipro.com; yashwanth.p54@wipro.com; niraj.kumar63@wipro.com; ayush.raj3@wipro.com; asif.ekbal@gmail.com

Abstract
Codemixing, the linguistic phenomenon where a speaker alternates between two or more languages within a conversation or even a single utterance, presents a significant challenge for machine translation systems due to its syntactic complexity and contextual nuances. This paper introduces a set of advanced transformer-based models fine-tuned specifically for translating codemixed text to English, more specifically, Hindi-English (colloquially referred to as "Hinglish") codemixed text into English. Unlike standard bilingual corpora, codemixed data requires an understanding of the intricacies of grammatical structures and cultural contexts embedded within the language blend. Existing machine translation efforts in codemixed languages have largely been constrained by the paucity of robust datasets and models that can capture the nuanced semantic and syntactic interplay characteristic of such languages. We present a novel dataset PACMAN for Hinglish to English machine translation, based on the PACMAN strategy, meticulously curated to represent natural codemixing patterns. Our generic fine-tuned translation models

trained on the novel data outperform current state-of-the-art Large Language Models (LLMs) by 38% in terms of BLEU score. Further, when fine-tuned on custom benchmark datasets, our focused dual fine-tuned models surpass the PHINC dataset BLEU score benchmark by 22%. Our comparative analysis illustrates significant improvements in translation quality, showcasing the potential of fine-tuning transformer models in bridging the linguistic divide in codemixed language translation. The success of our models reflects a promising step forward in the quest to provide seamless translation services for the ever-growing multilingual population and the complex linguistic phenomena they generate.

Automated System for Opinion Detection of Breathing Problem Discussions in Medical Forum Using Deep Neural Network

Somenath Nag Choudhury (IIT Patna); Asif Ekbal (Indian Institute of Technology, Patna)
somenath_2121cs27@iitp.ac.in; asif@iitp.ac.in

Abstract
Chest X-ray radiology majorly focuses on diseases like consolidation, pneumothorax, pleural effusion, lung collapse, etc., causing breathing and circulation problems. A tendency to share such problems in the forums for an answer without revealing personal demographics is also very common. However, we have observed more visitors than authors, which leads to a very poor average reply per discussion (3 to 12 only), and also many left with no or late replies in the forums. To alleviate the process, and ease of acquiring the best replies from multiple discussions, we propose a supervised learning framework by automatic scrapping and annotation of breathing problem-related group discussions from the patient.info\footnote{\url{https://patient.info/forums}}forum and determine the associated sentiment of the most voted respondent post using Bi-LSTM. We assume the most voted reply is the most factual and experienced reply. We mainly scrapped and determined the sentiment of bronchiectasis, asthma, pneumonia, and respiratory disease-related posts. After filtering and augmentation, a total of 1,756 posts were used for training our Stacked Bi-LSTM model and achieved an overall accuracy of 90 %.

Effect of Pivot Language and Segment-Based Few-Shot Prompting for Cross-Domain Multi-Intent Identification in Low Resource Languages

Kathakali Mitra (BITS PILANI , Hyderabad Campus); Aditha Venkata Santosh Ashish (BITS PILANI , Hyderabad Campus); Soumya Teotia (BITS PILANI , Hyderabad Campus); Aruna Malapati (BITS PILANI , Hyderabad Campus)
p20220104@hyderabad.bits-pilani.ac.in; f20191435@hyderabad.bits-pilani.ac.in; f20202182@hyderabad.bits-pilani.ac.in; arunam@hyderabad.bits-pilani.ac.in

Abstract
NLU (Natural Language Understanding) has considerable difficulties in identifying multiple intentions across different domains in languages with limited resources. Our contributions involve utilizing pivot languages with similar semantics for NLU tasks, creating a vector database for efficient retrieval and indexing of language embeddings in high-resource languages for Retrieval Augmented Generation (RAG) in low-resource languages, and thoroughly investigating the effect of segment-based strategies on complex user utterances across multiple domains and intents in the development of a Chain of Thought Prompting (COT) combined with Retrieval Augmented Generation. The study investigated recursive approaches to identify the most effective zero-shot instances for segment-based prompting. A comparison analysis was conducted to compare the effectiveness of sentence-based prompting vs segment-based prompting across different domains and multiple intents. This research offers a promising avenue to address the formidable challenges of NLU in low-resource languages, with potential applications in conversational agents and dialogue systems and a broader impact on linguistic understanding and inclusivity. Keywords : Retrieval Augmented Generation (RAG) , Chain of Thought Prompting (COT) ,

Ada Embedding , GPT 4 , Chroma Vector Database , Embedding , High Resource Language (HRL) , Low Resource Language (LRL) , Large Language Models (LLMs) , Pre-Trained Language Models (PLMs)

Towards Large Language Model driven Reference-less Translation Evaluation for English and Indian Language

Vandan Mujadia (IIIT-H); Pruthwik Mishra (IIIT, Hyderabad); Arafat Ahsan (IIIT Hyderabad); Dipti M. Sharma (IIITH)
vmujadia@gmail.com; pruthwikmishra@gmail.com; arafat.ahsan@iiit.ac.in; dipti@iiit.ac.in

Abstract
With the primary focus on evaluating the effectiveness of large language models for automatic reference-less translation assessment, this work presents our experiments on mimicking human direct assessment to evaluate the quality of translations in English and Indian languages. We constructed a translation evaluation task where we performed zero-shot learning, in-context example-driven learning, and fine-tuning of large language models to provide a score out of 100, where 100 represents a perfect translation and 1 represents a poor translation. We compared the performance of our trained systems with existing methods such as COMET, BERT-Scorer, and LABSE, and found that the LLM-based evaluator (LLaMA-2-14B) achieves a comparable or higher overall correlation with human judgments for the considered Indian language pairs.

1-step Speech Understanding and Transcription Using CTC Loss

Karan Singla (SAIL, University of Southern California); Shahab Jalalvand (Interactions LLC); Yeon-Jun Kim (Interactions LLC); Andrej Ljolje (AT&T Labs - Research); Antonio Moreno Daniel (Interactions LLC); Srinivas Bangalore (Interactions Corp); Benjamin Stern (Interactions LLC)
ksingla025@gmail.com; shahabjld2@gmail.com; ykim@interactions.com; alj@research.att.com; amoreno@gmail.com; srini65@live.com; benjamin.j.stern@gmail.com

Abstract
Recent studies have made some progress in refining end-to-end (E2E) speech recognition encoders by applying Connectionist Temporal Classification (CTC) loss to enhance named entity recognition within transcriptions. However, these methods have been constrained by their exclusive use of the ASCII character set, allowing only a limited array of semantic labels. Our proposed solution extends the E2E automatic speech recognition (ASR) system's vocabulary by adding a set of unused placeholder symbols, conceptually akin to the tokens used in sequence modeling. These placeholders are then assigned to represent semantic tags and are integrated into the transcription process as distinct tokens. We demonstrate notable improvements in entity tagging, intent discernment, and transcription accuracy on the SLUE benchmark and yields results that are on par with those for the SLURP dataset. Additionally, we provide a visual analysis of the system's proficiency in accurately pinpointing meaningful tokens over time, illustrating the enhancement in transcription quality through the utilization of supplementary semantic tags.

Consolidating Strategies for Countering Hate Speech Using Persuasive Dialogues

Sougata Saha (State University of New York at Buffalo); Rohini Srihari (University at Buffalo, SUNY)
sougatas@buffalo.edu; rohini@buffalo.edu

Abstract
Hateful comments are prevalent on social media platforms. Although tools for automatically detecting, flagging, and blocking such false, offensive, and harmful content online have lately matured, such reactive and brute force methods alone provide short-term and superficial remedies while the perpetrators persist. With the public availability of large language models which can generate articulate synthetic and engaging content at scale, there are concerns about the rapid growth of dissemination of such malicious content on the web. There is now a need to focus on deeper, long-term solutions that involve engaging with the human perpetrator behind the source of the content to change their viewpoint or at least bring down the rhetoric using persuasive means. To do that, we propose defining and experimenting with controllable strategies for generating counter-arguments to hateful comments in online conversations. We experiment with controlling response generation using features based on (i) argument structure and reasoning-based Walton argument schemes, (ii) counter-argument speech acts, and (iii) human characteristics-based qualities such as Big-5 personality traits and human values. Using automatic and human evaluations, we determine the best combination of features that generate fluent, argumentative, and logically sound arguments for countering hate. We further share the developed computational models for automatically annotating text with such features, and a silver-standard annotated version of an existing hate speech dialog corpora.

Poster Presentations

Konkani ASR

Swapnil Fadte (Computer Science and Technology, Goa University); Gaurish Thakkar (University of Zagreb); Jyoti Pawar (Goa University, Goa)
swapnil.fadte@unigoa.ac.in; gthakkar@m.ffzg.hr; jdp@unigoa.ac.in

Abstract
Konkani is a resource-scarce language, mainly spoken on the west coast of India. The lack of resources directly impacts the development of language technology tools and services. Therefore, the development of digital resources is required to aid in the improvement of this situation. This paper describes the work on the Automatic Speech Recognition (ASR) System for Konkani language. We have created the ASR by fine-tuning the whisper-small ASR model with 100 hours of Konkani speech corpus data. The baseline model showed a word error rate (WER) of 17, which serves as evidence for the efficacy of the fine-tuning procedure in establishing ASR accuracy for Konkani language.

Query-Based Summarization and Sentiment Analysis for Indian Financial Text by leveraging Dense Passage Retriever, RoBERTa, and FinBERT

Numair Shaikh (Department of Computer Engineering, SCTR's Pune Institute of Computer Technology); JayeshPatil (Department of Computer Engineering, SCTR's Pune Institute of Computer Technology); SheetalSonawane (Department of Computer Engineering, SCTR's Pune Institute of Computer Technology)
numairsh77@gmail.com; patiljayeshsunil@gmail.com; sssonawane@pict.edu

Abstract
With the ever-expanding pool of information accessible on the Internet, it has become increasingly challenging for readers to sift through voluminous data and derive meaningful insights. This is particularly noteworthy and critical in the context of documents such as financial reports and large-scale media reports. In the realm of finance, documents are typically lengthy and comprise numerical values. This research delves into the extraction of insights through text summaries from financial data, based on the user's interests, and the identification of clues from these insights. This research presents a straightforward, all-encompassing framework for conducting query-based summarization of financial documents, as well as analyzing the sentiment of the summary. The system's performance is evaluated using benchmarked metrics, and it is compared to State-of-The-Art (SoTA) algorithms. Extensive experimentation indicates that the proposed system surpasses existing pre-trained language models.

Bias Detection Using Textual Representation of Multimedia Contents

Karthik L Nagar (Accenture); Aditya Mohan Singh (Accenture Technology Labs); Sowmya Rasipuram (Accenture Technology Labs); RoshniRamnani (Accenture Technology labs); MilindSavagaonkar (Accenture); AnutoshMaitra (Accenture)
karthik.l.nagar@accenture.com; aditya.mohan.singh@accenture.com; sowmya.rasipuram@accenture.com; roshni.r.ramnani@accenture.com; milind.savagaonkar@accenture.com; anutosh.maitra@accenture.com

Abstract
The presence of biased and prejudicial content in social media has become a pressing concern, given its

potential to inflict severe societal damage. Detecting and addressing such bias is imperative, as the rapid dissemination of skewed content can disrupt social harmony. Advanced deep learning models are now paving the way for the automatic detection of bias in multimedia content with human-like accuracy. This paper focuses on identifying social bias in social media images. Toward this, we curated a Social Bias Image Dataset (SBID), consisting of 300 bias/no-bias images. The images contain both textual and visual information. We scientifically annotated the dataset for four different categories of bias. Our methodology involves generating a textual representation of the image content leveraging state-of-the-art models of optical character recognition (OCR), image captioning, and character attribute extraction. Initially, we performed fine-tuning on a Bidirectional Encoder Representations from Transformers (BERT) network to classify bias and no-bias, as well as on a Bidirectional Auto-Regressive Transformer (BART) network for bias categorization, utilizing an extensive textual corpus. Further, these networks were fine-tuned on the image dataset built by us - SBID. The experimental findings presented herein underscore the effectiveness of these models in identifying various forms of bias in social media images. We will also demonstrate their capacity to discern both explicit and implicit bias.

Rating YouTube Videos through Sentiment Analysis: A Case Study in Bangla-English code-mixed Situation

ArunavaKar (University of Hyderabad); Angshuman Jana (IIT Guwahati)
arunavakaronly@gmail.com; arunavakaronly@gmail.com

Abstract
Opinion mining from social media and its application has become an important area of research in the last few years. YouTube is the largest online repository of videos and it provides a comment section to enable viewers to express their opinion. This comment section can be a good source for analysis which can potentially lead to the development of many critical applications. This paper introduces a novel technique to rate YouTube videos utilising a Bangla sentiment analysis model built on the state-of-the-art transformer architecture. Our application links viewers' sentiments to a useful rating system, effectively empowering the users to have a rating system for YouTube videos just like Amazon or Google Play. Bangla speakers are predominantly multilingual and users of multiple scripts. The YouTube comments made by Bangla speakers display a high degree of code-mixed content. This necessitates the use of a specialised sentiment-analysis model for Bangla-English code-mixed texts. Leveraging a diverse dataset sourced from YouTube comments, the model achieves 93% accuracy after training and fine-tuning. Then, the model is integrated into an application that calculates opinion scores and sentiment categories. The application's accuracy is validated through a study involving a group of 20 volunteers, showing impressive results. Furthermore, we highlight that it can be used to detect many related tasks like clickbait video detection, unsafe content detection, etc.

Annotated and Normalized Causal Relation Extraction Corpus for Improving Health Informatics

Samridhi Dev (Jawaharlal Nehru University); Aditi Sharan (Jawaharlal Nehru University)
samrid21_scs@jnu.ac.in; aditisharan@gmail.com

Abstract
In the ever-expanding landscape of biomedical research, development of new cancer drugs has increased the likelihood of adverse drug reactions (ADRs) that can affect patient outcomes. However, information about these ADRs is often buried in unstructured data, requiring the conversion of this data into a structured and labeled dataset to identify potential ADRs and associations between them, making the extraction of entities and the analysis of causal relations a pivotal task. Machine learning methods have been used to identify ADRs, but current literature has several gaps in coverage, superficial manual annotation, and a lack of a labeled ADR corpus specific to cancer and normalized entities. Current

datasets are generated manually on the abstracts, limiting their scope. To address these limitations, the paper presents an algorithm that automatically constructs, annotates, normalizes entities specific to cancer and identifies causal relationships among entities using linguistics and grammatical properties, MetaMap and UMLS tools enabling efficient information retrieval. A further knowledge graph was created for a case report to visualize the causal relationships.

T20NGD: Annotated corpus for news headlines classification in low resource language,Telugu.

CHINDUKURI MALLIKARJUNA (NIT-TIRUCHIRAPPALLI); SangeethaSivanesan (NIT-TIRUCHIRAPPALLI)
malli.chindukuri@gmail.com; sangeetha@nitt.edu

Abstract
News classification allows analysts and researchers to study trends over time. Based on classification, news platforms can provide readers with related articles. Many digital news platforms and apps use classification to offer personalized content for their users. While there are numerous resources accessible for news classification in various Indian languages, there is still a lack of extensive benchmark dataset specifically for the Telugu language. Our paper presents and describes the Telugu20news group dataset, where news has been collected from various online Telugu news channels. We describe in detail the accumulation and annotation of the proposed news headlines dataset. In addition, we conducted extensive experiments on our proposed news headlines dataset in order to deliver solid baselines for future work.

Advancing Class Diagram Extraction from Requirement Text: A Transformer-Based Approach

Shweta . (Assistant Professor, Department of CSE, LNMIIT Jaipur); Suyash Mittal (Department of CSE, LNMIIT Jaipur); Suryansh Chauhan (Department of CSE, LNMIIT Jaipur)
shweta.singh332@gmail.com; 22ucs215@lnmiit.ac.in; 22ucs214@lnmiit.ac.in

Abstract
The class diagram plays an important role in software development. As these diagrams are created using software requirement text, it helps to improve communication between the developers and the stakeholders. Thus, the automatic extraction of class diagrams enhances the speed of software development procedures. The research carried out in this direction mostly relies on rule-based methodologies and deep learning models. These methodologies have their drawbacks, such as the fact that large rule-based systems are complex to handle, whereas the word embeddings used in deep learning are context-independent. Thus, the presented work strives to extract the class diagram entities from the natural language text by employing a transformer-based model, as the embeddings generated by these models are context-dependent. The results have been compared with the existing procedure, and an ablation study has also been carried out to find out the relevance of each step in the extraction procedure. The analysis involved examining the true positive, false positive, and false negative rates for specific class diagram elements in separate case studies. As a result, an enhancement of 9–7% has been observed in the procedures used for extracting the resulting class diagrams.

L3Cube-IndicNews: News-based Short Text and Long Document Classification Datasets in Indic languages

Aishwarya Mirashi (Pune Institute of Computer Technology); Srushti Sonavane (Indian); Purva Lingayat (Pune Institute of Computer Technology); Tejas Padhiyar (PICT); Raviraj Joshi (Indian Institute of Technology Madras)
aishwaryamirashi@gmail.com; srushtisonavane@gmail.com; lingayatpurva7@gmail.com; tejaspad240@gmail.com; raviraj.j1991@gmail.com

Abstract
In this work, we introduce L3Cube-IndicNews, a multilingual text classification corpus aimed at curating a high-quality dataset for Indian regional languages, with a specific focus on news headlines and articles. We have centered our work on 10 prominent Indic languages, including Hindi, Bengali, Marathi, Telugu, Tamil, Gujarati, Kannada, Odia, Malayalam, and Punjabi. Each of these news datasets comprises 10 or more classes of news articles. L3Cube-IndicNews offers 3 distinct datasets tailored to handle different document lengths that are classified as: Short Headlines Classification (SHC) dataset containing the news headline and news category, Long Document Classification (LDC) dataset containing the whole news article and the news category, and Long Paragraph Classification (LPC) containing sub-articles of the news and the news category. We maintain consistent labeling across all 3 datasets for in-depth length-based analysis. We evaluate each of these Indic language datasets using 4 different models including monolingual BERT, multilingual Indic Sentence BERT (IndicSBERT), and IndicBERT. This research contributes significantly to expanding the pool of available text classification datasets and also makes it possible to develop topic classification models for Indian regional languages. This also serves as an excellent resource for cross-lingual analysis owing to the high overlap of labels among languages. The datasets and models are shared publicly at https://github.com/l3cube-pune/indic-nlp.

PoS to UPoS Conversion and Creation of UPoS Tagged Resources for Assamese Language

Kuwali Talukdar (Gauhati University); Prof. Shikhar Kumar Sarma (Gauhati University)
kuwalitalukdar@gmail.com; sks001@gmail.com

Abstract
This paper addresses the vital task of transitioning from traditional Part-of-Speech (PoS) tagging to Universal Part-of-Speech (UPoS) tagging within the linguistic context of the Assamese language. The paper outlines a comprehensive methodology for PoS to UPoS conversion and the creation of UPoS tagged resources, bridging the gap between localized linguistic analysis and universal standards. The significance of this work lies in its potential to enhance natural language processing and understanding for the Assamese language, contributing to broader multilingual applications. The paper details the data preparation and creation processes, annotation methods, and evaluation techniques, shedding light on the challenges and opportunities presented in the pursuit of linguistic universality. The contents of this research have implications for improving language technology in the Assamese language and can serve as a model for similar work in other regional languages. Mapping of standard PoStagset applicable for Indian languages to that of the primary categories of the UPoStagset is done with respect to the Assamese language lexical behaviour. Conversion of PoS tagged text corpus to UPoStaged corpus using this mapping, and then utilizing a Deep Learning based model trained on such a dataset to create a sizable UPoS tagged corpus, are presented in a structured flow. This paper is a step towards a more standardized, universal understanding of linguistic elements in a diverse and multilingual world.

Mitigating Abusive Comment Detection in Tamil Text: A Data Augmentation Approach with Transformer Model

Reshma Sheik (National Institute of Technology, Trichy); Raghavan Balanathan (National Institute of Technology, Trichy); S Jaya Nirmala (National Institute of Technology Tiruchirappalli)
rezmasheik@gmail.com; raghavanb21@gmail.com; sjaya@nitt.edu

Abstract
With the increasing number of users on social media platforms, the detection and categorization of abusive comments have become crucial, necessitating effective strategies to mitigate their impact on online discussions. However, the intricate and diverse nature of low-resource Indic languages presents a challenge in developing reliable detection methodologies. This research focuses on the task of classifying YouTube comments written in Tamil language into various categories. To achieve this, our research conducted experiments utilizing various multi-lingual transformer-based models along with data augmentation approaches involving back translation approaches and other pre-processing techniques. Our work provides valuable insights into the effectiveness of various preprocessing methods for this classification task. Our experiments showed that the Multilingual Representations for Indian Languages (MURIL) transformer model, coupled with round-trip translation and lexical replacement, yielded the most promising results, showcasing a significant improvement of over 15 units in macro F1-score compared to existing baselines. This contribution adds to the ongoing research to mitigate the adverse impact of abusive content on online platforms, emphasizing the utilization of diverse preprocessing strategies and state-of-the-art language models.

Dravidian Fake News Detection with Gradient Accumulation based Transformer Model

Eduri Raja (National Institute of Technology Silchar); Badal Soni (National Institute of Technology Silchar); Samir Kumar Borgohain (National Institute of Technology Silchar); Candy Lalrempuii (National Institute of Technology Silchar)
eduri_rs@cse.nits.ac.in; badal@cse.nits.ac.in; samir@cse.nits.ac.in; candy_rs@cse.nits.a.cin

Abstract
The proliferation of fake news poses a significant challenge in the digital era. Detecting false information, especially in non-English languages, is crucial to combating misinformation effectively. In this research, we introduce a novel approach for Dravidian fake news detection by harnessing the capabilities of the MuRIL transformer model, further enhanced by gradient accumulation techniques. Our study focuses on the Dravidian languages, a diverse group of languages spoken in South India, which are often underserved in natural language processing research. We optimize memory usage, stabilize training, and improve the model's overall performance by accumulating gradients over multiple batches. The proposed model exhibits promising results in terms of both accuracy and efficiency. Our findings underline the significance of adapting state-of-the-art techniques, such as MuRIL-based models and gradient accumulation, to non-English languages to address the pressing issue of fake news.

Automatic Speech Recognition System for Malasar Language using Multilingual Transfer Learning

Basil K. Raju (Kerala University of Digital Sciences Innovation and Technology); Leena G. Pillai (Kerala University of Digital Sciences Innovation and Technology); Kavya Manohar (Kerala University of Digital Sciences Innovation and Technology); Elizabeth Sherly (Dr)
basil.cs21@duk.ac.in; leena.g@duk.ac.in; kavya.manohar@duk.ac.in; sherly@iiitmk.ac.in

Abstract
This study pioneers the development of an automatic speech recognition (ASR) system for the Malasar language, an extremely low-resource ethnic language spoken by a tribal community in the Western Ghats of South India. Malasar is primarily an oral language which does not have a native script. Therefore, Malasar is often transcribed in Tamil script, a closely related major language. This work presents the first ever effort of leveraging the capabilities of multilingual transfer learning for recognisingmalasar speech. We fine-tune a pre-trained multilingual transformer model with Malasar speech data. In our endeavour to fine-tune this model using a Malasar speech corpus, we could successfully bring down the WER to 48.00% from 99.08% (zero shot baseline). This work demonstrates the efficacy of multilingual transfer learning in addressing the challenges of ASR for extremely low-resource languages, contributing to the preservation of their linguistic and cultural heritage.

Dy-poThon: A Bangla Sentence-Learning System for Children with Dyslexia

DipshikhaPodder (Indian Institute of Technology Kharagpur); manjirasinha (Assistant Professor, Indian Institute of Technology Kharagpur); TirthankarDasgupta (Tata Consultancy Services Ltd.); AnupamBasu (IIT Kharagpur)
dipshi.podder@gmail.com; manjira87@gmail.com; iamtirthankar@gmail.com; anupambas@gmail.com

Abstract
The number of assistive technologies available for dyslexia in Bangla is low and most of them do not use multisensory teaching methods. As a solution, a computer-based audio-visual system Dy-poThon is proposed to teach sentence reading in Bangla. It incorporates a multisensory teaching method through three activities, listening, reading, and writing, checks the reading and writing ability of the user and tracks the response time. A criteria-based evaluation was conducted with 28 special educators to evaluate Dy-poThon. Content, efficiency, ease of use and aesthetics are evaluated using a standardised questionnaire. The result suggests that Dy-poThon is useful for teaching Bangla sentence-reading

Mitigating Clickbait: An Approach to Spoiler Generation Using Multitask Learning

Sayantan Pal (State University of New York at Buffalo); Souvik Das (University at Buffalo); Rohini Srihari (University at Buffalo, SUNY)
spal5@buffalo.edu; souvikda@buffalo.edu; rohini@buffalo.edu

Abstract
This study introduces 'clickbait spoiling', a novel technique designed to detect, categorize, and generate spoilers as succinct text responses, countering the curiosity induced by clickbait content. By leveraging a multi-task learning framework, our model's generalization capabilities are significantly enhanced, effectively addressing the pervasive issue of clickbait. The crux of our research lies in generating appropriate spoilers, be it a phrase, an extended passage, or multiple, depending on the spoiler type required. Our methodology integrates two crucial techniques: a refined spoiler categorization method and a modified version of the Question Answering (QA) mechanism, incorporated within a multi-task learning paradigm for optimized spoiler extraction from context. Notably, we have included fine-tuning

methods for models capable of handling longer sequences to accommodate the generation of extended spoilers. This research highlights the potential of sophisticated text processing techniques in tackling the omnipresent issue of clickbait, promising an enhanced user experience in the digital realm.

Comparing DAE-based and MASS-based UNMT: Robustness to Word-Order Divergence in English-->Indic Language Pairs

Tamali Banerjee (IIT Bombay); Rudra Murthy (IBM India Research Limited); Pushpak Bhattacharyya (Indian Institute of Technology Bombay and Patna)
banerjeetamali6@gmail.com; rmurthyv@in.ibm.com; pushpakbh@gmail.com

Abstract
We test the robustness of state-of-the-art Unsupervised NMT (UNMT) approaches (i.e., MASS-based UNMT and DAE-based UNMT) to word-order divergence between source and target languages. We investigate this by comparing two models for each of the two approaches, i.e., (i) model trained on language pairs with different word-orders, and (ii) model trained on the same language pairs with source language re-ordered to match the word-order of the target language. Ideally, UNMT approaches that are robust to word-order divergence should exhibit no visible performance difference between the two configurations. Our study focuses on five English --> Indic language pairs (i.e., en-hi, en-bn, en-gu, en-kn, and en-ta) with SVO source word-order and SOV target word-order. Our findings show that DAE-based UNMT consistently outperforms MASS-based UNMT in translation accuracy for these language pairs. Bridging the word-order gap through reordering improves the accuracy of MASS-based UNMT models but does not improve DAE-based UNMT models. This suggests that DAE-based UNMT is more robust to word-order divergence.

MahaSQuAD : Bridging Linguistic Divides in Marathi Question-Answering

Ruturaj Ghatage (Pune Institute of Computer Technology); Aditya Ashutosh Kulkarni (Pune Institute of Computer Technology); Rajlaxmi Patil (Pune Institute of Computer Technology); Sharvi Endait (Pune Institute of Computer Technology); Raviraj Joshi (Indian Institute of Technology Madras)
ruturajghatage33@gmail.com; adityakulkarni2501@gmail.com; rajlaxmipatilsarita@gmail.com; sharviendait@gmail.com; raviraj.j1991@gmail.com

Abstract
Question-answering systems have revolutionized information retrieval, but linguistic and cultural boundaries limit their widespread accessibility. This research endeavors to bridge the gap of the absence of efficient QnA datasets in low-resource languages by translating the English Question Answering Dataset (SQuAD) using a robust data curation approach. We introduce MahaSQuAD, the first-ever full SQuAD dataset for the Indic language Marathi, consisting of 118,516 training, 11,873 validation, and 11,803 test samples. Challenges in maintaining context and handling linguistic nuances are addressed, ensuring accurate translations. Moreover, as a QnA dataset cannot be simply converted into any low-resource language using translation, we need a robust method to map the answer translation to its span in the translated passage. Hence, to address this challenge, we also present a generic approach for translating SQuAD into any low-resource language. Thus, we offer a scalable approach to bridge linguistic and cultural gaps present in low-resource languages, in the realm of question-answering systems. The datasets and models are shared publicly at https://github.com/l3cube-pune/MarathiNLP

CASM - Context and Something More in Lexical Simplification

Atharva Kumbhar (SCTR'S Pune Institute of Computer Technology); Sheetal Sonawane (SCTR'S Pune Institute of Computer Technology); Dipali Kadam (SCTR'S Pune Institute of Computer Technology); Prathamesh Mulay (Pune Institute Of Computer Technology)
computationallinguisticlab@gmail.com; sssonawane@pict.edu; ddkadam@pict.edu; prathumulay@gmail.com

Abstract
Lexical Simplification is a challenging task that aims to improve the readability of text for non-native people, people with dyslexia, and any linguistic impairments. It consists of 3 components: 1) Complex Word Identification 2) Substitute Generation 3) Substitute Ranking. Current methods use contextual information as a primary source in all three stages of the simplification pipeline. We argue that while context is an important measure, it alone is not sufficient in the process. In the complex word identification step, contextual information is inadequate, moreover, heavy feature engineering is required to use additional linguistic features. This paper presents a novel architecture for complex word identification that uses a pre-trained transformer model's information flow through its hidden layers as a feature representation that implicitly encodes all the features required for identification. We portray how database methods and masked language modeling can be complementary to one another in substitute generation and ranking process that is built on the foundational pillars of Simplicity, Grammatical and Semantic correctness, and context preservation. We show that our proposed model generalizes well and outperforms the current state-of-the-art on well-known datasets.

Improving the Evaluation of NLP Approaches for Scientific Text Annotation with Ontology Embedding-Based Semantic Similarity Metrics

Pratik Devkota (University of North Carolina at Greensboro); Somya D. Mohanty (United Health Care); Prashanti Manda (University of North Carolina at Greensboro)
p_devkota@uncg.edu; smohanty@unitedhealthgroup.com; p_manda@uncg.edu

Abstract
Ontologies are widely used to represent data in a variety of scientific domains, including biology, physics, and geography. Ontology curation and annotation, the process of reading scientific text and associating words and phrases with appropriate ontology concepts, is essential for this representation. Natural language processing (NLP) techniques powered by deep learning have recently become prominent in the task of ontology annotation. However, traditional metrics of accuracy, such as precision and recall, cannot be used to evaluate the accuracy of these methods, as their output is ontology concepts, not independent entities. Semantic similarity metrics offer the capability of estimating partial accuracy and have been used in recent work to evaluate NLP methods for ontology annotation. Here, we present robust semantic similarity metrics created through the use of ontology embeddings. We test our metrics using gold standard data pertaining to evolutionary biology created by scientists in the Phenoscape project and show that they outperform traditional semantic similarity metrics, offering a more robust and accurate assessment of NLP approaches designed for ontology annotation.

A Survey of using Large Language Models for Generating Infrastructure as Code

K Ganesh Srivatsa (International Institute of Information Technology, Hyderabad); SabyasachiMukhopadhyay (IIIT Hyderabad); Ganesh Katrapati (International Institute of Information Technology); Manish Shrivastava (International Institute of Information Technology Hyderabad)

kalahasti.ganesh@research.iiit.ac.in; sabyasachi.m@research.iiit.ac.in; ganesh.katrapati@research.iiit.ac.in; m.shrivastava@iiit.ac.in

Abstract

Infrastructure as Code (IaC) is a revolutionary approach which has gained significant prominence in the Industry. IaC manages and provisions IT infrastructure using machine-readable code by enabling automation, consistency across the environments, reproducibility, version control, error reduction and enhancement in scalability. However, IaC orchestration is often a painstaking effort which requires specialised skills as well as a lot of manual effort. Automation of IaC is a necessity in the present conditions of the Industry and in this survey, we study the feasibility of applying Large Language Models (LLM) to address this problem. LLMs are large neural network-based models which have demonstrated significant language processing abilities and shown to be capable of following a range of instructions within a broad scope. Recently, they have also been adapted for code understanding and generation tasks successfully, which makes them a promising choice for the automatic generation of IaC configurations. In this survey, we delve into the details of IaC, usage of IaC in different platforms, their challenges, LLMs in terms of code-generation aspects and the importance of LLMs in IaC along with our own experiments. Finally, we conclude by presenting the challenges in this area and highlighting the scope for future research.

First Attempt at Building Parallel Corpora for Machine Translation of Northeast India's Very Low-Resource Languages

Atnafu Lambebo Tonja (InstitutoPolitécnico Nacional (IPN), Centro de Investigaciónen Computación (CIC)); Melkamu Mersha (University of Colorado Colorado Springs); Ananya Kalita (University of Colorado); Olga Kolesnikova (Centro de Investigacionen Computacion del Instituto Politecnico Nacional); Jugal Kalita (University of Colorado)

atnafu.lambebo@wsu.edu.et; mmersha@uccs.edu; ananyawolf@gmail.com; kolesolga@gmail.com; jkalita@uccs.edu

Abstract

This paper presents the creation of initial bilingual corpora for thirteen very low-resource languages of India, all from Northeast India. It also presents the results of initial translation efforts in these languages. It creates the first-ever parallel corpora for these languages and provides initial benchmark neural machine translation results for these languages. We intend to extend these corpora to include a large number of low-resource Indian languages and integrate the effort with our prior work with African and American-Indian languages to create corpora covering a large number of languages from across the world.

"Kurosawa": A Script Writer's Assistant

Prerak Gandhi (Indian Institute of Technology Bombay); Vishal Pramanik (IIT Bombay); Pushpak Bhattacharyya (Indian Institute of Technology Bombay and Patna)
prerakgandhi13@gmail.com; vishalpramanik35@gmail.com; pushpakbh@gmail.com

Abstract
Storytelling is the lifeline of the entertainment industry - movies, TV shows, and stand-up comedies, all need stories. A good and gripping script is the lifeline of storytelling and demands creativity and resource investment. Good scriptwriters are rare to find and often work under severe time pressure. Consequently, entertainment media are actively looking for automation. In this paper, we present an AI-based script-writing workbench called KUROSAWA which addresses the tasks of plot generation and script generation. Plot generation aims to generate a coherent and creative plot (600-800 words) given a prompt (15-40 words). Script generation, on the other hand, generates a scene (200-500 words) in a screenplay format from a brief description (15-40 words). Kurosawa needs data to train. We use a 4-act structure of storytelling to annotate the plot dataset manually. We create a dataset of 1000 manually annotated plots and their corresponding prompts/storylines and a gold-standard dataset of 1000 scenes with four main elements - scene headings, action lines, dialogues, and character names - tagged individually. We fine-tune GPT-3 with the above datasets to generate plots and scenes. These plots and scenes are first evaluated and then used by the scriptwriters of a large and famous media platform ErosNow. We release the annotated datasets and the models trained on these datasets as a working benchmark for automatic movie plot and script generation.

Text-2-Wiki: Summarization and Template-driven Article Generation

Jayant Panwar (International Institute of Information Technology); Radhika Mamidi (International Institute of Information Technology)
jayant.panwar@research.iiit.ac.in; radhika.mamidi@iiit.ac.in

Abstract
Users on Wikipedia collaborate in a structured and organized manner to publish and update articles on numerous topics, which makes Wikipedia a very rich source of knowledge. English Wikipedia has the most amount of information available (more than 6.7 million articles); however, there are few good informative articles on Wikipedia in Indian languages. Hindi Wikipedia has approximately only 160k articles. The same article in Hindi can be vastly different from its English version and generally contains less information. This poses a problem for native Indian language speakers who are not proficient in English. Therefore, having the same amount of information in Indian Languages will help promote knowledge among those who are not well-versed in English. Publishing the articles manually, like the usual process in Global English Wikipedia, is a time-consuming process. To get the amount of information in native Indian languages up-to-speed with the amount of information in English, automating the whole article generation process is the best option. In this study, we present a stage-wise approach ranging from Data Collection to Summarization and Translation, and finally ending with Template Creation. This approach ensures the efficient generation of a large amount of content in Hindi Wikipedia in less time. With the help of this study, we were able to successfully generate more than a thousand articles in Hindi Wikipedia with ease.

Blind Leading the Blind: A Social-Media Analysis of the Tech Industry

Tanishq Chaudhary (International Institute of Information Technology Hyderabad); Pulak Malhotra (International Institute of Information Technology Hyderabad); Radhika Mamidi (International Institute of Information Technology Hyderabad); PonnurangamKumaraguru (International Institute of Information Technology Hyderabad)
tanishq.chaudhary@research.iiit.ac.in; pulak.malhotra@students.iiit.ac.in; radhika.mamidi@iiit.ac.in; pk.guru@iiit.ac.in

Abstract
Online social networks have changed the way we perceive careers. A standard screening process for employees now involves profile checks on LinkedIn, X, and other platforms, with any negative opinions scrutinized. Blind, an anonymous social networking platform, aims to satisfy this growing need for taboo workplace discourse. In this paper, for the first time, we present a large-scale empirical text-based analysis of the Blind platform. We acquire and release two novel datasets: 63k Blind Company Reviews and 767k Blind Posts, containing over seven years of industry data. Using these, we analyze the Blind network, study drivers of engagement, and obtain insights into the last eventful years, preceding, during, and post-COVID-19, accounting for the modern phenomena of work-from-home, return-to-office, and the layoffs surrounding the crisis. Finally, we leverage the unique richness of the Blind content and propose a novel content classification pipeline to automatically retrieve and annotate relevant career and industry content across other platforms. We achieve an accuracy of 99.25% for filtering out relevant content, 78.41% for fine-grained annotation, and 98.29% for opinion mining, demonstrating the high practicality of our software.

A Unified Multi task Learning Architecture for Hate Detection Leveraging User-based Information

Prashant Kapil (Indian Institute of Technology); Asif Ekbal (Indian Institute of Technology)
prashant.pcs17@iitp.ac.in; asif@iitp.ac.in

Abstract
Hate speech, offensive language, aggression, racism, sexism, and other abusive language is a common phenomenon in social media. There is a need for AI based intervention which can filter hate content at scale. Most existing hate speech detection solutions have utilized the features by treating each post as an isolated input instance for the classification. In this research, we investigate this challenging issue by introducing a unique model that improves hate speech identification for the English language by utilising intra-user and inter-user-based information. We also utilize the metadata available and combine it with other textual-based features to detect hate. The investigations used single-task learning (STL) and multi-task learning (MTL) paradigms using deep neural networks, such as convolution neural network (CNN), gated recurrent unit (GRU), bidirectional encoder representations from the transformer (BERT), and A Lite BERT (ALBERT). We use three benchmark datasets and conclude that combining certain user features with textual features gives significant improvements in macro-F1 and weighted-F1.

Mytho-Annotator: An Annotation tool for Indian Hindu Mythology

Apurba Paul (Jadavpur University); Anupam Mondal (Institute of Engineering and Management); Sainik Mahata (Jadavpur University); Srijan Seal (JISCE); Prasun Sarkar (JISCE); Dipankar Das (Jadavpur University)
apurba.saitech@gmail.com; link.anupam@gmail.com; sainik.mahata@gmail.com; srijanseal03@gmail.com; prasunsarkarpersonal@gmail.com; dipankar.dipnil2005@gmail.com

Abstract
Mythology is a collection of myths, especially one belonging to a particular religious or cultural tradition. We observed that an annotation tool is essential to identify important and complex information from any mythological texts or corpora. Additionally, obtaining high-quality annotated corpora for complex information extraction including labelled text segments is an expensive and time-consuming process. Hence, in this paper, we have designed and deployed an annotation tool for Hindu mythology which is presented as Mytho-Annotator. Its easy-to-use web-based text annotation tool is powered by Natural Language Processing (NLP). This tool primarily labels three different categories such as named entities, relationships, and event entities. This annotation tool offers a comprehensive and adaptable annotation paradigm.

Transformer-based Bengali Textual Emotion Recognition

Md. Atabuzzaman (Department of Computer Science, Virginia Tech); Maksuda Bilkis Baby (Hajee Mohammad Danesh Science and Technology University); Md Shajalal (Fraunhofer FIT)
atabuzzaman@vt.edu; maksuda.cse34@gmail.com; md.shajalal@fit.fraunhofer.de

Abstract
Emotion recognition for high-resource languages has progressed significantly. However, resource-constrained languages such as Bengali have not advanced notably due to the lack of large benchmark datasets. Besides this, the need for more Bengali language processing tools makes the emotion recognition task more challenging and complicated. Therefore, we developed the largest dataset in this paper, consisting of almost 12k Bengali texts with six basic emotions. Then, we conducted experiments on our dataset to establish the baseline performance applying machine learning, deep learning, and transformer-based models as emotion classifiers. The experimental results demonstrate that the models achieved promising performance in Bengali emotion recognition.

Citation-Based Summarization of Landmark Judgments

Purnima Bindal (University of Delhi); Vikas Kumar (University of Delhi); Vasudha Bhatnagar (University of Delhi); Parikshet Sirohi (University of Delhi); Ashwini Siwal (University of Delhi)
pbindal@cs.du.ac.in; vikas@cs.du.ac.in; vbhatnagar@cs.du.ac.in; parikshet.sirohi@gmail.com; asiwal@law.du.ac.in

Abstract
Landmark judgments are of prime importance in the Common Law System because of their exceptional jurisprudence and frequent references in other judgments. In this work, we leverage contextual references available in citing judgments to create an extractive summary of the target judgment. We evaluate the proposed algorithm on two datasets curated from the judgments of Indian Courts and find the results promising.

Aspect and Opinion Term Extraction Using Graph Attention Network

Abir Chakraborty (Microsoft)
abir.chakraborty@gmail.com

Abstract
In this work we investigate the capability of Graph Attention Network for extracting aspect and opinion terms. Aspect and opinion term extraction is posed as a token-level classification task akin to named entity recognition. We use the dependency tree of the input query as additional feature in a Graph Attention Network along with the token and part-of-speech features. We show that the dependency structure is a powerful feature that in the presence of a CRF layer substantially improves the performance and generates the best result on the commonly used datasets from SemEval 2014, 2015 and 2016. We experiment with additional layers like BiLSTM and Transformer in addition to the CRF layer. We also show that our approach works well in the presence of multiple aspects or sentiments in the same query and it is not necessary to modify the dependency tree based on a single aspect as was the original application for sentiment classification.

Abstractive Hindi Text Summarization: A Challenge in a Low-Resource Setting

Daisy Lal (Lancaster University); Paul Rayson (Lancaster University); Krishna Singh (IIIT Allahabad); Uma Shanker Tiwary (Indian Institute of Information Technology Allahabad)
d.m.lal@lancaster.ac.uk; p.rayson@lancaster.ac.uk; kpsingh@iiita.ac.in; ust@iiita.ac.in

Abstract
The Internet has led to a surge in text data in Indian languages; hence, text summarization tools have become essential for information retrieval. Due to a lack of data resources, prevailing summarizing systems in Indian languages have been primarily dependent on and derived from English text summarization approaches. Despite Hindi being the most widely spoken language in India, progress in Hindi summarization is being delayed due to the lack of proper labeled datasets. In this preliminary work we address two major challenges in abstractive Hindi text summarization: creating Hindi language summaries and assessing the efficacy of the produced summaries. Since transfer learning (TL) has shown to be effective in low-resource settings, in order to assess the effectiveness of TL-based approach for summarizing Hindi text, we perform a comparative analysis using three encoder-decoder models: attention-based (BASE), multi-level (MED), and TL-based model (RETRAIN). In relation to the second challenge, we introduce the ICE-H evaluation metric based on the ICE metric for assessing English language summaries. The Rouge and ICE-H metrics are used for evaluating the BASE, MED, and RETRAIN models. According to the Rouge results, the RETRAIN model produces slightly better abstracts than the BASE and MED models for 20k and 100k training samples. The ICE-H metric, on the other hand, produces inconclusive results, which may be attributed to the limitations of existing Hindi NLP resources, such as word embeddings and POS taggers.

Verb Categorisation for Hindi Word Problem Solving

Harshita Sharma (iiit.ac.in); Pruthwik Mishra (IIIT, Hyderabad); Dipti Sharma (IIIT, Hyderabad)
harshita.sharma@research.iiit.ac.in; pruthwikmishra@gmail.com; diptims@gmail.com

Abstract
Word problem Solving is a challenging NLP task that deals with solving mathematical problems described in natural language. Recently, there has been renewed interest in developing word problem solvers for Indian languages. As part of this paper, we have built a Hindi arithmetic word problem solver which makes use of verbs. Additionally, we have created verb categorization data for Hindi. Verbs are

very important for solving word problems with addition/subtraction operations as they help us identify the set of operations required to solve the word problems. We propose a rule-based solver that uses verb categorisation to identify operations in a word problem and generate answers for it. To perform verb categorisation, we explore several approaches and present a comparative study.

ReviewCraft : A Word2Vec Driven System Enhancing User-Written Reviews

Gaurav Sawant (Goa University); Pradnya Bhagat (Goa University); Jyoti Pawar (Goa University,
gauravrsawant1313@gmail.com; dcst.pradanya@unigoa.ac.in; jdp@unigoa.ac.in

Abstract
The significance of online product reviews has become indispensable for customers in making informed buying decisions, while e-commerce platforms use them to fine tune their recommender systems. However, since review writing is purely a voluntary process without any incentives, most customers opt out from writing reviews or write poor-quality ones. This lack of engagement poses credibility issues as fake or biased reviews can mislead buyers who rely on them for informed decision-making. To address this issue, this paper introduces a system that suggests product features and appropriate sentiment words to help users write informative product reviews in a structured manner. The system is based on Word2Vec model and Chi square test. The evaluation results demonstrates that the reviews with recommendations showed a 2 fold improvement both, in the quality of the features covered and correct usage of sentiment words, as well as a 19% improvement in overall usefulness compared to reviews without recommendations.

Intent Detection and Zero-shot Intent Classification for Chatbots

Sobha Lalitha Devi (AU-KBC Research Centre, Anna University); Pattabhi RK Rao (AU-KBC Research centre)
sobha@au-kbc.org; pattabhi@au-kbc.org

Abstract
In this paper we give in detail how seen and unseen intent is detected and classified. User intent detection has a critical role in dialogue systems. While analysing the intents it has been found that intents are diversely expressed and new variety of intents emerge continuously. Here we propose a capsule-based approach that classifies the intent and a zero-shot learning to identify the unseen intent. There are recently proposed methods on zero-shot classification which are implemented differently from ours. We have also developed an annotated corpus of free conversations in Tamil, the language we have used for intent classification and for our chatbot. Our proposed method on intent classification performs well.

Coreference Resolution Using AdapterFusion-based Multi-Task learning

Sobha Lalitha Devi (AU-KBC Research Centre, Anna University); Vijay Sundar Ram (AU-KBC, Anna University, Chennai); Pattabhi RK Rao (AU-KBC Research centre)
sobha@au-kbc.org; sundar@au-kbc.org; pattabhi@au-kbc.org

Abstract
End-to-end coreference resolution is the task of identifying the mentions in a text that refer to the same real world entity and grouping them into clusters. It is crucially required for natural language understanding tasks and other high-level NLP tasks. In this paper, we present an end-to-end architecture for neural coreference resolution using AdapterFusion, a new two stage learning algorithm that leverages knowledge from multiple tasks. First task is in identifying the mentions in the text and the second to determine the coreference clusters. In the first task we learn task specific parameters called adapters that encapsulate the task-specific information and then combine the adapters in a separate knowledge

composition step to identify the mentions and their clusters. We evaluated it using FIRE corpus for Malayalam and Tamil and we achieved state of art performance.

Transfer learning in low-resourced MT: An empirical study

Sainik Mahata (Jadavpur University); Dipanjan Saha (Jadavpur University); Dipankar Das (Jadavpur University); Sivaji Bandyopadhyay (JADAVPUR UNIVERSITY, NIT SILCHAR)
sainik.mahata@gmail.com; sahadipanjan6@gmail.com; dipankar.dipnil2005@gmail.com; sivaji.cse.ju@gmail.com

Abstract
Translation systems rely on a large and good-quality parallel corpus for producing reliable translations. However, obtaining such a corpus for low-resourced languages is a challenge. New research has shown that transfer learning can mitigate this issue by augmenting low-resourced MT systems with high-resourced ones. In this work, we explore two types of transfer learning techniques, namely, cross-lingual transfer learning and multilingual training, both with information augmentation, to examine the degree of performance improvement following the augmentation. Furthermore, we use languages of the same family (Romanic, in our case), to investigate the role of the shared linguistic property, in producing dependable translations.

Transformer-based Nepali Text-to-Speech

Ishan Dongol (Kathmandu University); Bal Krishna Bal (Department of Computer Science and Engineering, Kathmandu University, Nepal)
ishandongol@gmail.com; bal@ku.edu.np

Abstract
Research on Deep learning-based Text-to-Speech (TTS) systems has gained increasing popularity in low-resource languages as this approach is not only computationally robust but also has the capability to produce state-of-the-art results. However, these approaches are yet to be significantly explored for the Nepali language, primarily because of the lack of adequate size datasets and secondarily because of the relatively sophisticated computing resources they demand. This paper explores the FastPitch acoustic model with HiFi-GAN vocoder for the Nepali language. We trained the acoustic model with two datasets, OpenSLR and a dataset prepared jointly by the Information and Language Processing Research Lab (ILPRL) and the Nepal Association of the Blind (NAB), to be further referred to as the ILPRLNAB dataset. We achieved a Mean Opinion Score (MOS) of 3.70 and 3.40 respectively for the same model with different datasets. The synthesized speech produced by the model was found to be quite natural and of good quality.

Infusing Knowledge into Large Language Models with Contextual Prompts

Kinshuk Vasisht (University of Delhi); Balaji Ganesan (IBM Research, India); Vikas Kumar (University of Delhi); Vasudha Bhatnagar (University of Delhi)
kinshuk.mcs21@cs.du.ac.in; bganesa1@in.ibm.com; vikas@cs.du.ac.in; vbhatnagar@cs.du.ac.in

Abstract
Knowledge infusion is a promising method for enhancing Large Language Models for domain-specific NLP tasks than pre-training models over large data from scratch. These augmented LLMs typically depend on additional pre-training or knowledge prompts from an existing knowledge graph, which is impractical in many applications. In contrast, knowledge infusion directly from relevant documents is more generalisable and alleviates the need for structured knowledge graphs while also being useful for

entities that are usually not found in any knowledge graph. With this motivation, we propose a simple yet generalisable approach for knowledge infusion by generating prompts from the context in the input text. Our experiments show the effectiveness of our approach which we evaluate by probing the fine-tuned LLMs.

Can Big Models Help Diverse Languages? Investigating Large Pretrained Multilingual Models for Machine Translation of Indian Languages

Telem Joyson Singh (IIT Guwahati); Sanasam Ranbir Singh (Indian Institute of Technology Guwahati); Priyankoo Sarmah (Indian Institute of Technology Guwahati)
joyson0117@gmail.com; ranbir@iitg.ernet.in; priyankoo@iitg.ac.in

Abstract
Machine translation of Indian languages is challenging due to several factors, including linguistic diversity, limited parallel data, language divergence, and complex morphology. Recently, large pre-trained multilingual models have shown promise in improving translation quality. In this paper, we conduct a large-scale study on applying large pre-trained models for English-Indic machine translation through transfer learning across languages and domains. This study systematically evaluates the practical gains these models can provide and analyzes their capabilities for the translation of the Indian language by transfer learning. Specifically, we experiment with several models, including Meta's mBART, mBART-many-to-many, NLLB-200, M2M-100, and Google's MT5. These models are fine-tuned on small, high-quality English-Indic parallel data across languages and domains. Our findings show that adapting large pre-trained models to particular languages by fine-tuning improves translation quality across the Indic languages, even for languages unseen during pretraining. Domain adaptation through continued fine-tuning improves results. Our study provides insights into utilizing large pretrained models to address the distinct challenges of MT of Indian languages.

Revolutionizing Authentication: Harnessing Natural Language Understanding for Dynamic Password Generation and Verification

Akram Al-Rumaim (Goa University); Jyoti D. Pawar (Goa University)
akramalrumaim@gmail.com; jdp@unigoa.ac.in

Abstract
In our interconnected digital ecosystem, API security is paramount. Traditional static password systems once used for API authentication, face vulnerabilities to cyber threats. This paper explores Natural Language Understanding (NLU) as a tool for dynamic password solutions, achieving 50\% accuracy. It investigates GPT-2 for dynamic password generation and innovative NLU-based verification using a set of specific criteria and threshold adjustments. The study highlights NLU's benefits, challenges, and prospects in enhancing API security. This approach is a significant stride in safeguarding digital interfaces amidst evolving Cyber Security threats.

Leveraging Empathy, Distress, and Emotion for Accurate Personality Subtyping from Complex Human Textual Responses

Soumitra Ghosh (Fondazione Bruno Kessler (FBK), Italy); Tanisha Tiwari (Indian Institute of Technology Patna); Chetna Painkra (Indian institute of technology Patna); Gopendra Vikram Singh (IIT Patna); Asif Ekbal (Indian Institute of Technology Patna)
sghosh@fbk.eu; tanishatiwari5@gmail.com; chetnapaikra55@gmail.com; gopendra.99@gmail.com; asif.ekbal@gmail.com

Abstract
Automated personality subtyping is a crucial area of research with diverse applications in psychology, healthcare, and marketing. However, current studies face challenges such as insufficient data, noisy text data, and difficulty in capturing complex personality traits. To address these issues, including empathy, distress, and emotion as auxiliary tasks in automated personality subtyping may enhance accuracy and robustness. This study introduces a Multi-input Multi-task Framework for Personality, Empathy, Distress, and Emotion Detection (MultiPEDE). This framework harnesses the complementary information from empathy, distress, and emotion tasks (auxiliary tasks) to enhance the accuracy and generalizability of automated personality subtyping (the primary task). The model uses a novel deep-learning architecture that captures the interdependencies between these constructs, is end-to-end trainable, and does not rely on ensemble strategies, making it practical for real-world applications. Performance evaluation involves labeled examples of five personality traits, two classes each for personality, empathy, and distress detection, and seven classes for emotion detection. This approach has diverse applications, including mental health diagnosis, improving online services, and aiding job candidate selection.

A Baseline System for Khasi and Assamese Bidirectional NMT with Zero available Parallel Data : Dataset Creation and System Development

Kishore Kashyap (Department of Information Technology , Gauhati University); Kuwali Talukdar (Gauhati University); Mazida Ahmed (Department of Information Technology, Gauhati University); Parvez Boruah (Gauhati University)
kb.guwahati@gmail.com; kuwalitalukdar@gmail.com; 14mazida.ahmed@gmail.com; parvezaziz70@gmail.com

Abstract
In this work we have tried to build a baseline Neural Machine Translation system for Khasi and Assamese in both directions. Both the languages are considered as low-resourced Indic languages. As per the language family in concerned, Assamese is a language from Indo-Aryan family and Khasi belongs to the Mon-Khmer branch of the Austroasiatic language family. No prior work is done which investigate the performance of Neural Machine Translation for these two diverse low-resourced languages. It is also worth mentioning that no parallel corpus and test data is available for these two languages. The main contribution of this work is the creation of Khasi-Assamese parallel corpus and test set. Apart from this, we also created baseline systems in both directions for the said language pair. We got best bilingual evaluation understudy (BLEU) score of 2.78 for Khasi to Assamese translation direction and 5.51 for Assamese to Khasi translation direction. We then applied phrase table injection (phrase augmentation) technique and got new higher BLEU score of 5.01 and 7.28 for Khasi to Assamese and Assamese to Khasi translation direction respectively.

Parts of Speech (PoS) and Universal Parts of Speech (UPoS) Tagging: A Critical Review with Special Reference to Low Resource Languages

Kuwali Talukdar (Gauhati University); Prof.Shikhar Kumar Sarma (Gauhati University); Manash Pratim Bhuyan (Dibru College, Dibrugarh, Assam)
kuwalitalukdar@gmail.com; sks001@gmail.com; mpratim250@gmail.com

Abstract
Universal Parts of Speech (UPoS) tags are parts of 2 speech annotations used in Universal Dependencies. 3 Universal Dependency (UD) helps in developing 4 cross-linguistically consistent treebank annotations 5 for multiple languages with a common framework 6 and standard. For various Natural Language 7 Processing (NLP) tasks and research such as semantic 8 parsing, syntactic parsing as well as linguistic parsing, 9 UD treebanks are becoming increasingly important 10 resources. A lot of interest has been seen in adopting 11 UD and UPoS standards and resources for integrating 12 with various NLP techniques, including Machine 13 Translations, Question Answering, Sentiment 14 Analysis etc. Consequently, a wide variety of 15 Artificial Intelligence (AI) and NLP tools are being 16 created with UD and UPoS standards on board. Part 17 of Speech (PoS) tagging is one of the fundamental 18 NLP tasks, which labels a specific sentence or set of 19 words in a paragraph with lexical and grammatical 20 annotations, based on the context of the sentence. 21 Contemporary Machine Learning (ML) and Deep 22 Learning (DL) techniques require god quality tagged 23 resources for training potential tagger models. Low 24 resource languages face serious challenges here. This 25 paper discusses about the UPoS in UD and presents a 26 concise yet inclusive piece of literature regarding 27 UPoS, PoS, and various taggers for multiple 28 languages with special reference to various low 29 resource languages. Already adopted approaches and 30 models developed for different low resource 31 languages are included in this review, considering 32 representations from a wide variety of languages. 33 Also, the study offers a comprehensive classification 34 based on the well-known ML and DL techniques used 35 in the development of part-of-speech taggers. This 36 will serve as a ready-reference for understanding 37 nuances of PoS and UPoS tagging.

Neural Machine Translation for Assamese-Bodo, a Low Resourced Indian Language Pair

Kuwali Talukdar (Gauhati University); Prof. Shikhar Kumar Sarma (Gauhati University); Farha Naznin (Gauhati University); Kishore Kashyap (Department of Information Technology , Gauhati University); Mazida Ahmed (Department of Information Technology, Gauhati University); Parvez Boruah (Gauhati University)
kuwalitalukdar@gmail.com; sks001@gmail.com; farha.gu@gmail.com; kb.guwahati@gmail.com; 14mazida.ahmed@gmail.com; parvezaziz70@gmail.com

Abstract
Results have been reported in 2 various works related to low resource 3 languages, using Neural Machine 4 Translation (NMT), where size of parallel 5 dataset is relatively low. This work presents 6 the experiment of Machine Translation in 7 the low resource Indian language pair 8 Assamese-Bodo, with a relatively low 9 amount of parallel data. Tokenization of 10 raw data is done with IndicNLP tool. NMT 11 model is trained with preprocessed dataset, 12 and model performances have been 13 observed with varying hyper parameters. 14 Experiments have been completed with 15 Vocab Size 8000 and 16000. Significant 16 increase in BLEU score has been observed 17 in doubling the Vocab size. Also data size 18 increase has contributed to enhanced 19 overall performances. BLEU scores have 20 been recorded with training on a data set of 21 70000 parallel sentences, and the results are 22 compared with another round of training 23 with a data set enhanced with 11500 24 Wordnet parallel data. A gold standard test 25 data set of 500 sentence size has been used 26 for recording BLEU. First round reported 27 an overall BLEU of 4.0, with vocab size of 28 8000. With same vocab size, and Wordnet 29 enhanced dataset, BLEU score of 4.33 was 30 recorded. Significant increase of BLEU 31 score (6.94) has been observed

with vocab 32 size of 16000. Next round of experiment 33 was done with additional 7000 new data, 34 and filtering the entire dataset. New BLEU 35 recorded was 9.68, with 16000 vocab size. 36 Cross validation has also been designed and 37 performed with an experiment with 8-fold 38 data chunks prepared on 80K total dataset. 39 Impressive BLEU scores of (Fold-1 40 through fold-8) 18.12, 16.28, 18.90, 19.25, 41 19.60, 18.43, 16.28, and 7.70 have been 42 recorded. The 8th fold BLEU deviated from 43 the trend, might be because of non-44 homogeneous last fold data.

Attentive Fusion: A Transformer-based Approach to Multimodal Hate Speech Detection

Atanu Mandal (Jadavpur University); Gargi Roy (Optum Global Solutions Private Limited, Bengaluru, India); Amit Barman (Jadavpur University); Indranil Dutta (Jadavpur University); Sudip Naskar (Jadavpur University)

atanumandal0491@gmail.com; roygargi1997@gmail.com; amitbarman811@gmail.com; indranildutta.lnl@jadavpuruniversity.in; sudip.naskar@gmail.com

Abstract

With the recent surge and exponential growth of social media usage, scrutinizing social media content for the presence of any hateful content is of utmost importance. Researchers have been diligently working since the past decade on distinguishing between content that promotes hatred and content that does not. Traditionally, the main focus has been on analyzing textual content. However, recent research attempts have also commenced into the identification of audio-based content. Nevertheless, studies have shown that relying solely on audio or text-based content may be ineffective, as recent upsurge indicates that individuals often employ sarcasm in their speech and writing. To overcome these challenges, we present an approach to identify whether a speech promotes hate or not utilizing both audio and textual representations. Our methodology is based on the Transformer framework that incorporates both audio and text sampling, accompanied by our very own layer called "Attentive Fusion"'. The results of our study surpassed previous state-of-the-art techniques, achieving an impressive macro F1 score of 0.927 on the Test Set.

Handwritten Text Segmentation Using U-Net and Shuffled Frog-Leaping Algorithm with Scale Space Technique

MoumitaMoitra (NIT Durgapur); Sujan Kumar Saha (National Institute of Technology Durgapur)

mou.moitra12@gmail.com; sujan.kr.saha@gmail.com

Abstract

The paper introduces a new method for segmenting words from handwritten Bangla documents. We found that the available handwritten character recognition (HCR) systems do not provide the desired accuracy in recognizing the text written by school students. Recognizing students' handwritten text becomes challenging due to certain factors, including a non-uniform gap between lines and words, and ambiguous, overlapping characters. The performance may be improved if the words in the text are segmented correctly before recognition. For the segmentation, we propose a combination of U-Net and a modified Scale Space method enhanced by the Shuffled Frog-Leaping Algorithm (SFLA). We employ the U-Net model for line segmentation; it effectively handles the variable spacing and skewed lines. After line segmentation, for segmenting the words, we use SFLA with Scale Space, allowing adaptive scaling and optimized parameter tuning. The proposed technique has been tested on two datasets: the openly available BN-HTR dataset and an in-house dataset prepared by collecting Bengali handwritten answer books from schools. In our experiments, we found that the proposed technique achieved promising performance on both datasets.

Identifying Correlation between Sentiment Analysis and Septic News Sentences Classification Tasks

Soma Das (Indian Institute of Information Technology Kalyani); Sagarika Ghosh (Indian Institute Of Information Technology Kalyani); Sanjay Chatterji (Indian Institute of Information Technology Kalyani)
soma_phd_2018july@iiitkalyani.ac.in; sagarika_phd_2018july@iiitkalyani.ac.in; sanjayc@iiitkalyani.ac.in

Abstract
This research investigates the correlation between Sentiment and SEPSIS(SpEculation, oPinion, biaS, and twISt) characteristics in news sentences through an ablation study. Various Sentiment analysis models, including TextBlob, Vader, and RoBERTa, are examined to discern their impact on news sentences. Additionally, we explore the Logistic Regression(LR), Decision Trees(DT), Support Vector Machines(SVM) and Convolutional Neural Network (CNN) models for Septic sentence classification.

KT2: Kannada-Tulu Parallel Corpus Construction for Neural Machine Translation

Asha Hegde (Mangalore University); Hosahalli Lakshmaiah Shashirekha (Mangalore University)
hegdekasha@gmail.com; hlsrekha@gmail.com

Abstract
In the last decade, Neural Machine Translation (NMT) has experienced substantial advancements. However, its widespread success has revealed a limitation in its performance when dealing with under-resourced language pairs, mainly due to the lack of parallel corpora in comparison to high-resourced language pairs like English-German, English-Spanish, and English-French. As a result, researchers have increasingly focused on implementing NMT techniques tailored to under-resourced language pairs and thereby the construction/collection of parallel corpora. In this view, this paper outlines the practical strategies for building a Kannada-Tulu parallel corpus and explores Machine Translation (MT) between these two Dravidian languages. Both Kannada and Tulu are under-resourced due to lack of processing tools and digital resources, especially parallel corpora, which are critical for MT development. Hence, this paper describes Kannada-Tulu parallel corpus construction in two folds: i) Manual Translation and ii) Automatic Text Generation (ATG). This work investigates various NMT approaches, including Recurrent Neural Networks (RNN), Bidirectional RNN (BiRNN), and transformer-based architectures, trained with Gated Recurrent Units (GRU) and Long Short Term Memory (LSTM) units. Additionally, the study explores sub-word tokenization techniques for translating Kannada to Tulu (Kn-Tu) and Tulu to Kannada (Tu-Kn) language pairs. The performance of these NMT models is evaluated using Character n-gram F-score (CHRF) and Bilingual Evaluation Understudy (BLEU) scores.

Enhancing Telugu Part-of-Speech Tagging with Deep Sequential Models and Multilingual Embeddings

Sai Rishith Reddy Mangamuru (Amrita School of Artificial Intelligence, Coimbatore, Amrita Vishwa Vidyapeetham, India); Sai Prashanth Karnati (Amrita School of Artificial Intelligence, Coimbatore, Amrita Vishwa Vidyapeetham, India); Bala Karthikeya Sajja (Amrita School of Artificial Intelligence, Coimbatore, Amrita Vishwa Vidyapeetham, India); Divith Phogat (Amrita School of Artificial Intelligence, Coimbatore, Amrita Vishwa Vidyapeetham, India); Premjith B (Amrita School of Artificial Intelligence, Coimbatore, Amrita Vishwa Vidyapeetham, India)
rishithmangamuru@gmail.com; karnatisaiprashanth@gmail.com; balakarthikeya19@gmail.com; divithphogatgp@gmail.com; prem.jb@gmail.com

Abstract
Part-of-speech (POS) tagging is a fundamental task in natural language processing (NLP) that involves assigning grammatical categories to words in a sentence. In this study, we investigate the application of deep sequential models for POS tagging of Telugu, a low-resource Dravidian language with rich morphology. We use the Universal dependencies dataset for this research and explore various deep learning architectures, including Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) networks, Gated Recurrent Units (GRUs), and their stacked variants for POS tagging. Additionally, we utilize multilingual BERT embeddings and indicBERTembeddings to capture contextual information from the input sequences. Our experiments demonstrate that stacked LSTM with multilingual BERT embeddings achieves the highest performance, outperforming other approaches and attaining an F1 score of 0.8812. These findings suggest that deep sequential models, particularly stacked LSTMs with multilingual BERT embeddings, are effective tools for POS tagging in Telugu.

Unlocking Emotions in Text: A Fusion of Word Embeddings and Lexical Knowledge for Emotion Classification

Anjali Bhardwaj (South Asian University, New Delhi, India); Nesar Ahmad Wasi (South Asian University, New Delhi, India); Muhammad Abulaish (South Asian University)
bhardwaj.anjali200594@gmail.com; nesarahmadwasi17@gmail.com; abulaish@ieee.org

Abstract
This paper introduces an improved method for emotion classification through the integration of emotion lexicons and pre-trained word embeddings. The proposed method utilizes semantically similar features to reconcile the semantic gap between words and emotions. The proposed approach is compared against three baselines for predicting Ekman's emotions at the document level on the GoEmotions dataset. The effectiveness of the proposed approach is assessed using standard evaluation metrics, which show at least a 5% gain in performance over baselines.

Convolutional Neural Networks can achieve binary bail judgement classification

Amit Barman (Jadavpur University); Devangan Roy (Jadavpur University); Debapriya Paul (Indian Institute of Engineering Science and Technology, Shibpur); Indranil Dutta (Jadavpur University); Shouvik Kumar Guha (Assistant Professor (Law), WBNUJS); Samir Karmakar(Jadavpur University); Sudip Naskar (Jadavpur University)
amitbarman811@gmail.com; devanganr.sll.rs@jadavpuruniversity.in; debapriya.rs2023@cs.iiests.ac.in; indranildutta.lnl@jadavpuruniversity.in; shouvikkumarguha@gmail.com; samir.karmakar@jadavpuruniversity.in; sudip.naskar@gmail.com

Abstract
There is an evident lack of implementation of Machine Learning (ML) in the legal domain in India, and any research that does take place in this domain is usually based on data from the higher courts of law and works with English data. The lower courts and data from the different regional languages of India are often overlooked. In this paper, we deploy a Convolutional Neural Network (CNN) architecture on a corpus of Hindi legal documents. We perform a bail Prediction task with the help of a CNN model and achieve an overall accuracy of 93% which is an improvement on the benchmark accuracy, set by Kapoor et al. (2022), albeit in data from 20 districts of the Indian state of Uttar Pradesh.

Multiset Dual Summarization for Incongruent News Article Detection

Sujit Kumar (Research Scholar, Indian Institute of Technology Guwahati); Rohan Jaiswal (Indian Institute of Technology Guwahati); Mohit Ram Sharma (Indian Institute of Technology Guwahati); Sanasam Ranbir Singh (Indian Institute of Technology Guwahati)
sujitkumar@iitg.ac.in; therohanjaiswal@gmail.com; mohit.rs@iitg.ac.in; ranbir@iitg.ac.in

Abstract
The prevalence of deceptive and incongruent news headlines has highlighted their substantial role in the propagation of fake news, exacerbating the spread of both misinformation and disinformation. Existing studies on incongruity detection primarily concentrate on estimating the similarity between the encoded representation of headlines and the encoded representation or summary representative vector of the news body. In the process of obtaining the encoded representation of the news body, researchers often consider either sequential encoding or hierarchical encoding of the news body or to acquire a summary representative vector of the news body, they explore techniques like summarization or dual summarization methods. Nevertheless, when it comes to detecting partially incongruent news, dual summarization-based methods tend to outperform hierarchical encoding-based methods. On the other hand, for datasets focused on detecting fake news, where the hierarchical structure within a news article plays a crucial role, hierarchical encoding-based methods tend to perform better than summarization-based methods. Recognizing this contradictory performance of hierarchical encoding-based and summarization-based methods across datasets with different characteristics, we introduced a novel approach called "Multiset Dual Summarization" (MDS). MDS combines the strengths of both hierarchical encoding and dual summarization methods to leverage their respective advantages. We conducted experiments on datasets with diverse characteristics, and our findings demonstrate that our proposed model outperforms established state-of-the-art baseline models.

Word Sense Disambiguation for Marathi language using Supervised Learning

Rasika Ransing (DattaMeghe College of Engineering, Vidyalankar Institute of Technology); Archana Gulati (School of Business Management, NMIMS University, Mumbai)
rasikaransing275@gmail.com; archana.gulati@nmims.edu

Abstract
The task of disambiguating word senses, often referred to as Word Sense Disambiguation (WSD), is a substantial difficulty in the realm of natural language processing. Marathi is widely acknowledged as a language that has a relatively restricted range of resources. Consequently, there has been a paucity of academic research undertaken on the Marathi language. There has been little research conducted on supervised learning for Marathi Word Sense Disambiguation (WSD) mostly owing to the scarcity of sense-annotated corpora. This work aims to construct a sense-annotated corpus for the Marathi language and further use supervised learning classifiers, such as Naïve Bayes, Support Vector Machine, Random Forest, and Logistic Regression, to disambiguate polysemous words in Marathi. The performance of these classifiers is evaluated.

A comparative study of transformer and transfer learning MT models for English-Manipuri

Kshetrimayum Boynao Singh (National Institute of Technology Silchar); Ningthoujam Avichandra Singh (National Institute of Technology Silchar); Loitongbam Sanayai Meetei (National Institute of Technology Silchar); Ningthoujam Justwant Singh (National Institute Of Technology, Silchar); Thoudam Doren Singh (National Institute of Technology Silchar); Sivaji Bandyopadhyay (JADAVPUR UNIVERSITY, NIT SILCHAR)
boynfrancis@gmail.com; avichandra0420@gmail.com; loisanayai@gmail.com; njustwant92@gmail.com; thoudam.doren@gmail.com; sivaji.cse.ju@gmail.com

Abstract
In this work, we focus on the development of machine translation (MT) models of a low-resource language pair viz. English-Manipuri. Manipuri is one of the eight scheduled languages of the Indian constitution. Manipuri is currently written in two different scripts: one is its original script called Meitei Mayek and the other is the Bengali script. We evaluate the performance of English-Manipuri MT models based on transformer and transfer learning technique. Our MT models are trained using a dataset of 69,065 parallel sentences and validated on 500 sentences. Using 500 test sentences, the English to Manipuri MT models achieved a BLEU score of 19.13 and 29.05 with mT5 and OpenNMT respectively. The results demonstrate that the OpenNMT model significantly outperforms the mT5 model. Additionally, Manipuri to English MT system trained with OpenNMT model reported a BLEU score of 30.90. We also carried out a comparative analysis between the Bengali script and the transliterated Meitei Mayek script for English-Manipuri MT models. This analysis reveals that the transliterated version enhances the MT model performance resulting in a notable +2.35 improvement in the BLEU score.

The Current Landscape of Multimodal Summarization

Atharva Kumbhar (SCTR'S Pune Institute of Computer Technology); Harsh Kulkarni (SCTR'S Pune Institute of Computer Technology); Atmaja Mali (SCTR'S Pune Institute of Computer Technology); Sheetal Sonawane (SCTR'S Pune Institute of Computer Technology); Prathamesh Mulay (SCTR'S Pune Institute of Computer Technology)
computationallinguisticlab@gmail.com; harshkulkarni1105@gmail.com; atmajamali07@gmail.com; sssonawane@pict.edu; prathumulay@gmail.com

Abstract
In recent years, the rise of multimedia content on the internet has inundated users with a vast and diverse array of information, including images, videos, and textual data. Handling this flood of multimedia data necessitates advanced techniques capable of distilling this wealth of information into concise, meaningful summaries. Multimodal summarization, which involves generating summaries from multiple modalities such as text, images, and videos, has become a pivotal area of research in natural language processing, computer vision, and multimedia analysis. This survey paper offers an overview of the state-of-the-art techniques, methodologies, and challenges in the domain of multimodal summarization. We highlight the interdisciplinary advancements made in this field specifically on the lines of two main frontiers:1) Multimodal Abstractive Summarization, and 2) Pre-training Language Models in Multimodal Summarization. By synthesizing insights from existing research, we aim to provide a holistic understanding of multimodal summarization techniques.

Automated Answer Validation using Text Similarity

Balaji Ganesan (IBM Research); Arjun Ravikumar (INDIAN INSTITUTE OF SCIENCE); Lakshay Piplani (Independent); Rini Bhaumik ; Dhivya Padmanaban (Indian Institute of Science); Shwetha Narasimhamurthy (Independent Researcher); Chetan Adhikary (Tata Consultancy Services); Subhash Deshapogu (Independent Researcher)
bganesa1@in.ibm.com; ark.arjun97@gmail.com; lakshay.piplani94@gmail.com; rini.bhaumik@gmail.com; dhivyathamil@gmail.com; shwethahybris@gmail.com; chetanadhikary@gmail.com; subhashdeshapogu@gmail.com

Abstract
Automated answer validation can help improve learning outcomes by providing appropriate feedback to learners, and by making question answering systems and online learning solutions more widely available. There have been some works in science question answering which show that information retrieval methods outperform neural methods, especially in the multiple choice version of this problem. We implement Siamese neural network models and produce a generalised solution to this problem. We compare our supervised model with other text similarity based solutions.

QeMMA: Quantum-Enhanced Multi-Modal Sentiment Analysis

Arpan Phukan (Indian Institute of Technology Patna); Asif Ekbal (Indian Institute of Technology Patna)
arpanphukan@gmail.com; asif.ekbal@gmail.com

Abstract
Multi-modal data analysis presents formidable challenges, as developing effective methods to capture correlations among different modalities remains an ongoing pursuit. In this study, we address multi-modal sentiment analysis through a novel quantum perspective. We propose that quantum principles, such as superposition, entanglement, and interference, offer a more comprehensive framework for capturing not

only the cross-modal interactions between text, acoustics, and visuals but also the intricate relations within each modality. To empirically evaluate our approach, we employ the CMU-MOSEI dataset as our testbed and utilize Qiskit by IBM to run our experiments on a quantum computer. Our proposed Quantum-Enhanced Multi-Modal Analysis Framework (QeMMA) showcases its significant potential by surpassing the baseline by 3.52% and 10.14% in terms of accuracy and F1 score, respectively, highlighting the promise of quantum-inspired methodologies.

Automatic Data Retrieval for Cross Lingual Summarization

Nikhilesh Bhatnagar (International Institute of Information Technology, Hyderabad); Ashok Urlana (TCS Research); Pruthwik Mishra (IIIT, Hyderabad); Vandan Mujadia (IIIT-H); Dipti Sharma (IIIT, Hyderabad)

tingc9@gmail.com; ashok.urlana@tcs.com; pruthwikmishra@gmail.com; vmujadia@gmail.com; diptims@gmail.com

Abstract

Cross lingual summarization involves the summarization of text written in one language to a different one. There is a body of research addressing cross-lingual summarization from English to other European languages. In this work, we aim to perform cross-lingual summarization from English to Hindi. We propose pairing up the coverage of newsworthy events in textual and video format can prove to be helpful for data acquisition for cross lingual summarization. We analyze the data and propose methods to match articles to video descriptions that serve as document and summary pairs. We also outline filtering methods over reasonable thresholds to ensure correctness of the summaries. Further, we make available 28,583 mono and cross-lingual article-summary pairs. We also build and analyze multiple baselines on the collected data and report error analysis.

Cross-Lingual Fact Checking: Automated Extraction and Verification of Information from Wikipedia using References

Shivansh Subramanian (IIIT Hyderabad); Ankita Maity (IIIT Hyderabad); Aakash Jain (IIIT Hyderabad); Bhavyajeet Singh (IIIT Hyderabad); Harshit Gupta (International Institute of Information Technology, Hyderabad); Lakshya Khanna (IIIT Hyderabad); Vasudeva Varma (IIIT Hyderabad)

shivansh.s@research.iiit.ac.in; ankita.maity@research.iiit.ac.in; aakash.jain@students.iiit.ac.in; bhavyajeet.singh@research.iiit.ac.in; harshit.g@research.iiit.ac.in; lakshya.khanna@research.iiit.ac.in; vv@iiit.ac.in

Abstract

The paper presents a novel approach for automated cross-lingual fact-checking that extracts and verifies information from Wikipedia using references. The problem involves determining whether a factoid in an article is supported or needs additional citations based on the provided references, with granularity at the fact level. We introduce a cross-lingual manually annotated dataset for fact extraction and verification and an entirely automated pipeline for the task. The proposed solution operates entirely in a cross-lingual setting, where the article text and the references can be in any language. The pipeline integrates several natural language processing techniques to extract the relevant facts from the input sources. The extracted facts are then verified against the references, leveraging the semantic relationships between the facts and the reference sources. Experimental evaluation on a large-scale dataset demonstrates the effectiveness and efficiency of the proposed approach in handling cross-lingual fact-checking tasks. We make our code and data publicly available.

Combining Pretrained Speech and Text Encoders for Continuous Spoken Language Processing

Karan Singla (SAIL, University of Southern California); Mahnoosh Mehrabani (Interactions LLC); Daniel Pressel (member); Ryan Price (Interactions); Bhargav Srinivas Chinnari (Interactions LLC); Yeon-Jun Kim (Interactions LLC); Srinivas Bangalore (Interactions Corp)
ksingla025@gmail.com; mahnoosh@interactions.net; dpressel@interactions.com; rprice@interactions.com; baru.anits@gmail.com; ykim@interactions.com; srini65@live.com

Abstract
In this paper, we propose a novel architecture for multi-modal speech and text input. We combine pretrained speech and text encoders using multi-headed cross-modal attention and jointly fine-tune on the target problem. The resultant architecture can be used for continuous token-level classification or utterance-level prediction acting on simultaneous text and speech. The resultant encoder efficiently captures both acoustic-prosodic and lexical information. We compare the benefits of multi-headed attention-based fusion for multi-modal utterance-level classification against a simple concatenation of pre-pooled, modality-specific representations. Our model architecture is compact, resource efficient, and can be trained on a single consumer GPU card.