SemEval 2024

The 18th International Workshop on Semantic Evaluation (SemEval-2024)

Proceedings of the Workshop

June 20-21, 2024

The SemEval organizers gratefully acknowledge the support from the following organizations.

Gold

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
209 N. Eighth Street
Stroudsburg, PA 18360
USA
Tel: +1-570-476-8006
Fax: +1-570-476-0860
acl@aclweb.org

ISBN 979-8-89176-107-0

Introduction

The Semantic Evaluation (SemEval) workshops focus on the evaluation and comparison of systems that analyze diverse semantic phenomena in text, with the aim of extending the current state of the art in semantic analysis and creating high quality annotated datasets in a range of increasingly challenging problems in natural language semantics. SemEval provides an exciting forum for researchers to propose challenging research problems in semantics and to build systems/techniques to address such research problems.

SemEval-2024 is the eighteenth workshop in the series of International Workshops on Semantic Evaluation. The workshop began in 1998 and was originally known as SensEval and focused on word sense disambiguation.
In 2007, the workshop was renamed SemEval, and evolved to include semantic tasks beyond word sense disambiguation. Starting in 2012, SemEval has been organized every year. The tasks for the next iteration of the workshop, SemEval-2025 (https://semeval.github.io/SemEval2025/), are underway.

SemEval-2024 is co-located (hybrid) with the 2024 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2024). SemEval-2024 will be held in Mexico City, Mexico and it includes the following 10 tasks:

- Semantic Relations
  - Task 1: Semantic Textual Relatedness for African and Asian Languages
  - Task 2: Safe Biomedical Natural Language Inference for Clinical Trials
- Discourse and Argumentation
  - Task 3: The Competition of Multimodal Emotion Cause Analysis in Conversations
  - Task 4: Multilingual Detection of Persuasion Techniques in Memes
  - Task 5: Argument Reasoning in Civil Procedure
- LLM Capabilities
  - Task 6: SHROOM, a Shared-task on Hallucinations and Related Observable Overgeneration Mistakes
  - Task 7: NumEval: Numeral-Aware Language Understanding and Generation
  - Task 8: Multidomain, Multimodel and Multilingual Machine-Generated Text Detection
- Knowledge Representation and Reasoning
  - Task 9: BRAINTEASER: A Novel Task Defying Common Sense
  - Task 10: Emotion Discovery and Reasoning its Flip in Conversation

This volume contains both the task description papers (10), that describe each of the above tasks, and the system description papers (279) that present the systems that participated in the tasks.
In addition, SemEval-2024 features two awards, one for the organizers of a task and one for a team participating in a task. The Best Task award recognizes a task that stands out for making an important intellectual contribution to empirical computational semantics, as demonstrated by a creative, interesting, and scientifically rigorous dataset and evaluation design, and a well-written task overview paper. The three Best System Description Paper awards recognize a system description paper (written by a team participating in one of the tasks) that advances our understanding of a problem and available solutions with respect to a task. It does not need to be the highest scoring system in the task, but it should have a strong analysis component in the evaluation, as well as a clear and reproducible description of the problem, algorithms, and methodology.

We are grateful to the task organizers for their dedication in carrying out ten very successful tasks and to the large number of participants whose enthusiastic participation has made SemEval 2024 a successful event. We also appreciate the efforts of the task organizers and participants who reviewed the paper submissions. These proceedings have greatly benefited from their detailed and thoughtful feedback. Finally, we also thank the members of the program committee who reviewed the submitted task proposals and helped us to select this exciting set of tasks, the NAACL 2024 conference organizers for their support, and the ACL Special Interest Group on the Lexicon (SIGLEX) for sponsoring and supporting this event.

Atul Kr. Ojha, A. Seza Doğruöz, Harish Tayyar Madabushi, Giovanni Da San Martino, Sara Rosenthal, and Aiala Rosá (SemEval-2024 Organizers and Co-Chairs)

Organizing Committee

SemEval Chairs

Atul Kr. Ojha, Data Science Institute, Unit for Linguistic Data, University of Galway
A. Seza Doğruöz, Universiteit Gent
Harish Tayyar Madabushi, University of Bath
Giovanni Da San Martino, University of Padova
Sara Rosenthal, IBM Research
Aiala Rosá, Instituto de Computación, Facultad de Ingeniería, Universidad de la República

Program Committee

Alessandro Raganato, University of Milano-Bicocca
Alexandra Uma, Queen Mary University of London
Ali Zeynali, College of Information and Computer Science, University of Massachusetts at Amherst
Amitava Das, University of of South Carolina
Andrey Kutuzov, University of Oslo
Anjie Fang, Amazon
Besnik Fetahu, Amazon
Boyuan Zheng, Ohio State University, Columbus
Carla Perez-Almendros, Cardiff University
Chenxi Whitehouse, City, University of London
Chico Q Camargo, University of Exeter
Chung-chi Chen, National Institute of Advanced Industrial Science and Technology
Corey A. Harper, Elsevier Labs
Cristiano Chesi, Istituto Universitario di Studi Superiori
Cunxiang Wang, Westlake University
David Jurgens, University of Michigan - Ann Arbor
Dimitar Dimitrov, University of Sofia "St. Kliment Ohridski"
Egoitz Laparra, University of Arizona
Elaine Zosa, SiloGen
Fanfan Wang, Nanjing University of Science and Technology
Filip Ilievski, Vrije Universiteit Amsterdam
Firoj Alam, Qatar Computing Research Institute, HBKU
Francesca Gasparini, University of Milano - Bicocca
Francesco Barbieri, Snap Inc.
Gavin Abercrombie, Heriot-Watt University
Gustavo H. Paetzold, Universidade Tecnologica Federal do Parana
Henning Wachsmuth, Leibniz Universitat Hannover
Ibrahim Abu Farha, University of Sheffield
Ibrahim Said Ahmad, Northeastern University
Idris Abdulmumin, Ahmadu Bello University, Zaria
Ivan Habernal, Paderborn University
Jakub Piskorski, Institute of Computer Science, Polish Academy of Science
Jeffrey S. Sorensen, Google
Jennifer D'Souza, TIB Hannover
Jeremy Barnes, University of the Basque Country
Jianfei Yu, Nanjing University of Science and Technology
Jingxuan Tu, Brandeis University

John Pavlopoulos, Stockholm University
Jonibek Mansurov, MBZUAI
Jose Camacho-Collados, Cardiff University
Kaixin Ma, Carnegie Mellon University
Lena Held, Technische UniversitÃ¤t Darmstadt
Mael Jullien, university of Manchester
Maik Frobe, Martin-Luther Universitat Halle-Wittenberg
Marcos Garcia, Universidade de Santiago de Compostela
Marko Robnik-Sikonja, University of Ljubljana
Matthew Purver, Queen Mary University of London
Matthew Shardlow, The Manchester Metropolitan University
Maximilian Heinrich, Bauhaus Universitat Weimar
Md. Shad Akhtar, Indraprastha Institute of Information Technology, Delhi
Meriem Beloucif, Uppsala University
Mohamed Abdalla, University of Toronto
Mohammad Taher Pilehvar, Tehran Institute for Advanced Studies
Nailia Mirzakhmedova, Bauhaus Universitat Weimar
Nedjma Ousidhoum, Cardiff University
Nicolas Stefanovitch, European Commission
Osama Mohammed Afzal, MBZUAI
Paolo Rosso, Universitat Politecnica de Valencia
Preslav Nakov, Mohamed bin Zayed University of Artificial Intelligence
Raul Vazquez, University of Helsinki
Roberto Zamparelli, University of Trento
Rui Xia, Nanjing University of Science and Technology
Saif Mohammad, NRC
Sanchit Ahuja, Microsoft Research
Seid Muhie Yimam, Universität Hamburg
Shervin Malmasi, Amazon
Shivani Kumar, Indraprastha Institute of Information Technology
Somin Wadhwa, Northeastern University
Steven Bethard, University of Arizona
Steven Schockaert, Cardiff University
Tanmoy Chakraborty, IIT Delhi
Teemu Vahtola, University of Helsinki
Timothee Mickus, University of Helsinki
Timothy A Miller, Harvard University
Tommaso Fornaciari, Italian National Police
Vivek Khetan, Accenture Labs
Vladimir Araujo, KU Leuven
Xi Chen, Department of Computer Science, University of Massachusetts at Amherst
Yifan Jiang, Information Sciences Institute,University of Southern California
Yuxia Wang, The University of Melbourne
Zhiyu Chen, Amazon

Keynote Talk: Beyond Single Scores: Transparent Evaluation through Fine-Grained Error Detection and Uncertainty Quantification

André F. T. Martins
Instituto Superior Tecnico, Senior Researcher at the Instituto de Telecomunicacoes, and VP of AI Research at Unbabel in Lisbon, Portugal

Abstract: Automatic evaluation metrics are key to drive progress in NLP. We use them to compare systems and decide which models to deploy, to understand the strengths and weaknesses of each model, and to help practitioners overcome existing failure modes. In this talk, I will discuss evaluation of machine translation quality. Today, lexical-based metrics (such as BLEU or ChrF) are being replaced by learned neural-based metrics, such as COMET and BLEURT, which exhibit much better correlation with human judgments. However, these metrics provide a single sentence-level score, offering little insight into translation errors (e.g., what are the errors and what is their severity). Can we do better? I will start by presenting xCOMET, an open-source learned metric which integrates both sentence-level evaluation and error span detection capabilities, exhibiting state-of-the-art performance across all types of evaluation (sentence-level, system-level, and error span detection). Moreover, it does so while highlighting and categorizing error spans, thus enriching the quality assessment. Then, I will discuss recent approaches that endow evaluation metrics with uncertainty quantification capabilities, using techniques such as Monte Carlo dropout, deep ensembles, heteroscedastic regression, quantile regression, and conformal prediction. Finally, I will present Tower, an open multilingual LLM for translation-related tasks. We perform continued pretraining on a multilingual mixture of monolingual and parallel data, creating TowerBase, followed by finetuning on instructions relevant for translation processes, creating TowerInstruct. The final model surpasses open alternatives on several tasks relevant to translation workflows and is competitive with general-purpose closed LLMs. To facilitate future research, we release the Tower models, our specialization dataset, an evaluation framework for LLMs focusing on the translation ecosystem, and a collection of model generations, including ours, on our benchmark.

Bio: André F. T. Martins is an Associate Professor at Instituto Superior Técnico, Senior Researcher at the Instituto de Telecomunicações, and VP of AI Research at Unbabel in Lisbon, Portugal. I also do scientific consulting for Priberam Labs. I work on natural language processing and machine learning.

Until 2012, André was a PhD student in the joint CMU-Portugal program in Language Technologies, at Carnegie Mellon University and Instituto Superior Técnico. His advisors were Mario Figueiredo, Noah Smith, Pedro Aguiar and Eric Xing.

Table of Contents

Program

Thursday, June 20, 2024

09:10 - 09:25 *Welcome and Introduction to SemEval*

09:30 - 10:30 *Invited Talk 1, Shared with *SEM*

10:30 - 11:00 *Coffee Break*

11:00 - 12:30 *Oral Session-I*

SemEval Task 1: Semantic Textual Relatedness for African and Asian Languages
Nedjma Ousidhoum, Shamsuddeen Hassan Muhammad, Mohamed Abdalla, Idris Abdulmumin, Ibrahim Said Ahmad, Sanchit Ahuja, Alham Fikri Aji, Vladimir Araujo, Meriem Beloucif and Christine De Kock

SemEval-2024 Task 2: Safe Biomedical Natural Language Inference for Clinical Trials
Mael Jullien, Marco Valentino and André Freitas

SemEval-2024 Task 3: Multimodal Emotion Cause Analysis in Conversations
Fanfan Wang, Heqing Ma, Rui Xia, Jianfei Yu and Erik Cambria

SemEval-2024 Task 4: Multilingual Detection of Persuasion Techniques in Memes
Dimitar Dimitrov, Firoj Alam, Maram Hasanain, Abul Hasnat, Fabrizio Silvestri, Preslav Nakov and Giovanni Da San Martino

SemEval-2024 Task 5: Argument Reasoning in Civil Procedure
Lena Held and Ivan Habernal

SemEval-2024 Task 6: SHROOM, a Shared-task on Hallucinations and Related Observable Overgeneration Mistakes
Timothee Mickus, Elaine Zosa, Raul Vazquez, Teemu Vahtola, Jörg Tiedemann, Vincent Segonne, Alessandro Raganato and Marianna Apidianaki

12:30 - 14:00 *Lunch*

14:00 - 15:00 *Oral Session-II*

SemEval-2024 Task 7: Numeral-Aware Language Understanding and Generation
Chung-chi Chen, Jian-tao Huang, Hen-hsen Huang, Hiroya Takamura and Hsin-hsi Chen

Thursday, June 20, 2024 (continued)

SemEval-2024 Task 8: Multidomain, Multimodel and Multilingual Machine-Generated Text Detection
Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Osama Mohammed Afzal, Tarek Mahmoud, Giovanni Puccetti and Thomas Arnold

SemEval-2024 Task 9: BRAINTEASER: A Novel Task Defying Common Sense
Yifan Jiang, Filip Ilievski and Kaixin Ma

SemEval 2024 - Task 10: Emotion Discovery and Reasoning its Flip in Conversation (EDiReF)
Shivani Kumar, Md. Shad Akhtar, Erik Cambria and Tanmoy Chakraborty

15:00 - 15:30 *Best System Paper's Presentations*

15:30 - 16:00 *Coffee Break*

16:00 - 17:30 *Poster Session I: System Description Papers (local and online)*

Friday, June 21, 2024

09:30 - 10:30 *Invited Talk: Beyond Single Scores: Transparent Evaluation through Fine-Grained Error Detection and Uncertainty Quantification (André F. T. Martins)*

10:30 - 11:00 *Coffee Break*

11:00 - 12:30 *Poster Session II: System Description Papers (online)*

12:30 - 14:00 *Lunch*

14:00 - 15:30 *Poster Session III: System Description Papers (in presence)*

15:30 - 16:00 *Coffee Break*

16:00 - 16:45 *Best Paper Awards and Concluding Remarks*