Eval4NLP 2022

Evaluation and Comparison of NLP Systems

Proceedings of the Third Workshop

November 20, 2022

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 978-1-959429-00-5

Introduction

Welcome to the Third Workshop on Evaluation and Comparison of NLP Systems (Eval4NLP 2022).

Fair evaluations and comparisons are essential for tracking development and identifying issues of NLP systems. In particular, recent NLP research has become increasingly dependent on fine-tuning pre-trained language models to perform downstream tasks, which has resulted in a considerable increase in the number of published state-of-the-art results. Such findings would be meaningless or even detrimental to the community without appropriate evaluation of all research aspects, including, but not limited to methodologies, datasets, metrics, and setups. To address these challenges, the Eval4NLP workshop series takes a broad and unifying perspective on the subject matter. The third edition of Eval4NLP workshop collocated with AACL 2022 continues to offer a forum for showcasing and discussing the most recent developments in NLP evaluation methods and resources.

Our workshop has attracted a lot of attention from the community with 20 research papers being submitted. After thorough consideration by the program committee and the workshop organizers, 11 papers were selected for presentation. This year's program covers a variety of topics in NLP evaluation and comparison, including new evaluation metrics (e.g., resource-performance tradeoff, summarization); systematic analyses over existing NLP models and techniques (e.g., GPT-2, stance classification baselines, data augmentation); new benchmark datasets for tasks like word segmentation, part-of-speech tagging, chat translation error detection, and multilingual referring expression generation; and critical analyses over existing evaluation benchmarks (e.g., STS) and metrics (e.g., SMATCH); and a novel adversarial example generation method.

We would like to thank all of the authors for their contributions, the program committee for their thoughtful reviews, the keynote speakers for sharing their perspectives, and all the attendees for their participation. We believe that all of these will contribute to a lively and successful workshop. Looking forward to meeting you all (virtually) at Eval4NLP 2022!

Eval4NLP 2022 Organization Team,
Daniel Deutsch, Can Udomcharoenchaikit, Juri Opitz,Yang Gao, Marina Fomicheva, Steffen Eger

Organizing Committee

Daniel Deutsch, Google Research, United States
Can Udomcharoenchaikit, Vidyasirimedhi Institute of Science and Technology, Thailand
Juri Opitz, Heidelberg University, Germany
Yang Gao, Google Research, United Kingdom
Marina Fomicheva, University of Sheffield, United Kingdom
Steffen Eger, Bielefeld University, Germany

Program Committee

Timothy Baldwin, MBZUAI and The University of Melbourne
Gerard De Melo, Hasso Plattner Institute and University of Potsdam
Daniel Deutsch, Google Research
Li Dong, Microsoft Research
Zi-Yi Dou, University of California, Los Angeles
Rotem Dror, School of Engineering and Applied Science, University of Pennsylvania
Steffen Eger, Bielefeld University
Ori Ernst, Bar-Ilan University
George Foster, Google
Anette Frank, Ruprecht-Karls-Universität Heidelberg
Yang Gao, Google Research
Yunsu Kim, Pohang University of Science and Technology
Lucy H. Lin, Spotify
Nitika Mathur, Oracle
Juri Opitz, Heidelberg University
Ines Rehbein, Universität Mannheim
Ehud Reiter, University of Aberdeen
Leonardo F. R. Ribeiro, Amazon Alexa AI
Ori Shapira, Amazon
Julius Steen, Institute for Computational Linguistics, Heidelberg University
Can Udomcharoenchaikit, Vidyasirimedhi Institute of Science and Technology
Shiyue Zhang, The University of North Carolina at Chapel Hill

Keynote Talk: SMART: Sentences as Basic Units for Text Evaluation

Reinald Kim Amplayo
Google

Abstract: Widely used evaluation metrics for text generation do not work well with longer multi-sentence texts. In this talk, I will introduce a new metric called SMART to mitigate such limitations. SMART treats sentences as basic units of matching instead of tokens, and uses a sentence matching function to soft-match candidate and reference sentences. Candidate sentences are also compared to sentences in the source documents to allow grounding (e.g., factuality) evaluation. Results show that system-level correlations of our proposed metric with a model-based matching function outperforms all competing metrics on the SummEval summarization meta-evaluation dataset, while the same metric with a string-based matching function is competitive with current model-based metrics. The latter does not use any neural model, which is useful during model development phases where resources can be limited and fast evaluation is required. SMART also outperforms all factuality evaluation metrics on the TRUE benchmark. Finally, extensive analyses show that our proposed metrics work well with longer summaries and are less biased towards specific models.

Bio: Reinald is a research scientist at Google working on text generation. Prior to that, he was a PhD student at the University of Edinburgh working with Mirella Lapata on opinion summarization. He was also affiliated with Yonsei University and Ateneo de Davao University.

Keynote Talk: Questioning Implicit Assumptions in our Evaluation Methodologies

Maxime Peyrard
EPFL

Abstract: Research in NLP/ML is driven by evaluation results, with attention and resources being focused on methods identified as state-of-the-art. The proper design of evaluation methodologies is thus crucial to ensure progress in the field. In this talk, we will discuss and review several assumptions implicitly made by our standard evaluation methodology and show that these assumptions may not be justified and have a significant impact on which systems are promoted to SotA.

Bio: Maxime Peyrard is a Post-Doc at EPFL in the data science lab. He is working at the intersection between NLP, and data science with a particular focus on methodological aspects like "how to obtain valid causal answers from data?" and "how to properly evaluate machine learning models?"

Table of Contents

Program

Sunday, November 20, 2022

10:30 - 10:45 *Opening Presentation*

11:30 - 12:15 *Paper Presentation Session 1*

Why is sentence similarity benchmark not predictive of application-oriented task performance?
Kaori Abe, Sho Yokoi, Tomoyuki Kajiwara and Kentaro Inui

Better Smatch = Better Parser? AMR evaluation is not so simple anymore
Juri Opitz and Anette Frank

Chat Translation Error Detection for Assisting Cross-lingual Communications
Yunmeng Li, Jun Suzuki, Makoto Morishita, Kaori Abe, Ryoko Tokuhisa, Ana Brassard and Kentaro Inui

13:15 - 14:00 *SMART: Sentences as Basic Units for Text Evaluation (Keynote Talk by Reinald Kim Amplayo)*

14:00 - 15:00 *Paper Presentation Session 2*

A Japanese Corpus of Many Specialized Domains for Word Segmentation and Part-of-Speech Tagging
Shohei Higashiyama, Masao Ideuchi, Masao Utiyama, Yoshiaki Oida and Eiichiro Sumita

Evaluating the role of non-lexical markers in GPT-2's language modeling behavior
Roberta Rocca and Alejandro de la Vega

From COMET to COMES – Can Summary Evaluation Benefit from Translation Evaluation?
Mateusz Krubiński and Pavel Pecina

Random Text Perturbations Work, but not Always
Zhengxiang Wang

15:30 - 16:15 *Questioning Implicit Assumptions in our Evaluation Methodologies (Keynote Talk by Maxime Peyrard)*

16:15 - 17:15 *Paper Presentation Session 3*