INLG 2024

The 17th International Natural Language Generation Conference

Tutorial Abstract

September 24, 2024

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 979-8-89176-127-8

Preface

This tutorial has been a long time in the making, at least in the sense that the work that has fed into it spans the last fifteen years or more. Over the last few years, calls for shared best practice and standardisation (including from us) have increased in frequency and urgency, so this seemed an opportune moment to collect recommendations into a single set of resources for ease of reference.

The resulting Tutorial on Human Evaluation of NLP System Quality is intended as a one-stop repository of information, guidance, and tools for putting together human evaluations in NLP. On the tutorial GitHub page, we make available papers, slides, lecture recordings, code, interactive tools and examples, intended to cover every one of the four phases of developing a human evaluation: Design, Implementation, Execution, and Analysis of Results.

The tutorial is conceived as a living repository where we collect feedback and suggestions for additions and improvements which will feed into periodic updates including for future live editions of the tutorial to be delivered at other conferences and/or summer schools.

The content of the tutorial has been given some of its shape by our collaborations with other researchers over the years which we gratefully acknowledge, in particular: Mohammad Arvan, Dave Howcroft, Emiel van Miltenburg, Natalie Parde, Maja Popovic, Ehud Reiter, and Anastasia Shimorina.

Lastly, we would like to thank the INLG 2024 Programme Chairs, Workshop Chair, Publication Chair and Local Organisers for their support in making this tutorial happen.

Anya Belz, João Sedoc, Craig Thomson, Simon Mille, Rudali Huidrom

Programme

Time	Duration (mins)	Unit #	Topic
09:30-10:00	30	Unit 1	Introduction
10:00-10:30	30	Unit 2	Development and Components of Human Evaluations
10:30-10:45	15	Break	
10:45-11:45	60	Unit 3	Quality Criteria and Evaluation Modes
11:45-12:30	45	Unit 4	Experiment Design
12:30-14:00	90	Lunch	
14:00-15:15	75	Unit 5	Statistical Analysis of Results
15:15-15:30	15	Break	
15:30-16:15	45	Unit 6	Experiment Implementation
16:15-16:40	25	Unit 7	Experiment Execution
16:40-16:55	15	Break	
16:55-18:30	95	Unit 8	Practical Session

Organizers

Tutorial Organizers

Anya Belz, ADAPT Research Centre, Dublin City University, Ireland
João Sedoc, New York University, USA
Craig Thomson, ADAPT Research Centre, Dublin City University, Ireland
Simon Mille, ADAPT Research Centre, Dublin City University, Ireland
Rudali Huidrom, ADAPT Research Centre, Dublin City University, Ireland

Local Organization Committee

Tatsuya Ishigaki (lead), National Institute of Advanced Industrial Science and Technology
Ayana Niwa, , Recruit Co., Ltd. / Megagon Labs
Takashi Yamamura, Yamagata Universit
Shun Tanaka, JX PRESS Corporation
Yumi Hamazano, Hitachi, Ltd.
Toshiki Kawamoto, Amazon
Takato Yamazaki LY Corp. / SB Intuitions Corp.
Hiroya Takamura, National Institute of Advanced Industrial Science and Technology
Ichiro Kobayashi, Kobayashi

Publication Chair

Chung-Chi Chen (National Institute of Advanced Industrial Science and Technology, Japan)

Table of Contents

The INLG 2024 Tutorial on Human Evaluation of NLP System Quality: Background, Overall Aims, and Summaries of Taught Units
Anya Belz, João Sedoc, Craig Thomson, Simon Mille and Rudali Huidrom 1