Human Evaluation of NLP System Quality

INLG Tutorial, 24th September 2024

Unit 1: Introduction

Link to Unit 1 Resources

Overview

Unit 1: Introduction

1. Unit aims, learning outcomes, contents
2. What is human evaluation in NLP?
3. Background
4. Challenges and issues
5. Overview of tutorial
6. Unit summary
7. References

Overview

Unit 1: Introduction

1. Unit aims, learning outcomes, contents
2. What is human evaluation in NLP?
3. Background
4. Challenges and issues
5. Overview of tutorial
6. Unit summary
7. References

Unit aims and learning outcomes

- The aims of Unit 1 are:
  - To give a first idea what human evaluation means in NLP.
  - To summarise the current state of human evaluation in NLP.
  - To survey some of the challenges and issues that have been identified, and how current research is beginning to address them.
  - To clarify the scope of the tutorial and present an overview of its units.
- After completion of the unit, participants will have basic knowledge of:
  - Foundational concepts and current state of human evaluation in NLP.
  - Current research topics, issues and challenges in human evaluation in NLP.
  - The scope and content of this tutorial.

Prerequisites and connections with other units

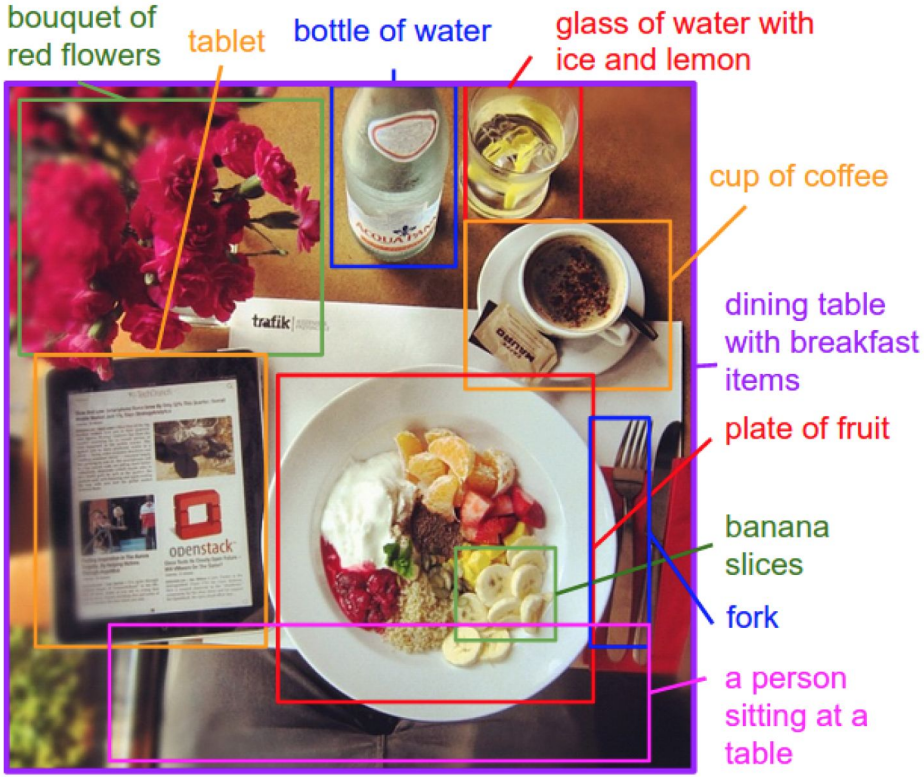
- No prior experience with, or knowledge of, human evaluation is required.
- The tutorial assumes basic familiarity with the field of Natural Language Processing.
- Unit 1 is not strictly a prerequisite for the other units, but introduces basic concepts that will be used in them, and provides an overview of their contents.

Overview

Unit 1: Introduction

1. Unit aims, learning outcomes, contents
2. What is human evaluation of NLP?
3. Background
4. Challenges and issues
5. Overview of tutorial
6. Unit summary
7. References

What is human evaluation in NLP?



- Example NLP system: image caption generator that takes a tagged image as input, and outputs a joined up caption for the image.
- Some possible captions are:

  Dining table with breakfast items.

  Dining table with breakfast items, including bouquet of red flowers, tablet, bottle of water, glass of water with ice and lemon, cup of coffee, plate of fruit, banana slices, fork, and a person sitting at a table.

  My dream breakfast.

  Where's the bacon and eggs?!

- Which of these captions are good? Better? On what grounds?

What is human evaluation in NLP?

Some ways in which NLP answers these questions:

- Take a dataset of example input/output pairs and for each, measure the similarity between system outputs and example outputs: BLEU, METEOR, ROUGE, chrF++, BERTscore, etc.
- Take some automatically computable measure of outputs, e.g. the diversity of the tokens in them: Self-BLEU, compression ratio, n-gram diversity, etc.

metric evaluation

- Ask some humans to rate how good each output is in terms of a given criterion: Fluency, Grammaticality, Input Coverage, etc.
- Ask some humans to perform a task with/without the outputs, and measure relative performance: number of post-edits, speed of finding searched-for items, etc.
- Ask some humans to interact with the system, and take automatically computable measurements during the interaction: task completion, click rates, reaction time, etc.
- Ask some humans to interact with the system, and ask them questions about their experience afterwards: overall satisfaction, ease of use, understandability, etc.

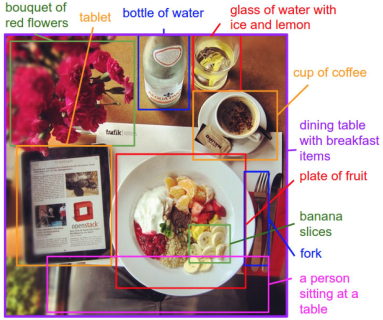
human evaluation

- Ask an LLM how good the outputs are given the inputs.

'LLM as judge'

What is human evaluation in NLP?

- Looking at the input image, intuitively might say that which caption is best depends on the context: is this image alt-text? Is it for accessibility?
- If it's for accessibility then the second caption could be a good candidate.
- If it's for social media, the last two might be good.
- But which of the last two is better depends on the poster's (dis)like of the image content.
- The seven different evaluation approaches above will give us different views of system quality.
- To get a fully rounded view of system quality, usually a combination of different evaluation methods is needed that are not strongly correlated with each other overall.



Dining table with breakfast items.

Dining table with breakfast items, including bouquet of red flowers, tablet, bottle of water, glass of water with ice and lemon, cup of coffee, plate of fruit, banana slices, fork, and a person sitting at a table.

My dream breakfast.

Where's the bacon and eggs?!

What is human evaluation in NLP?

- For each method applied, care needs to be taken that everything impacting the given view of quality is taken into account, e.g. application context and user perspective above.
- Over the course of the tutorial, will look at how to put together experiments that produce reliable answers to those questions (*Which of these captions are good? Better? On what grounds?*)

Overview

Unit 1: Introduction

1. Unit aims, learning outcomes, contents
2. What is human evaluation in NLP?
3. Background
4. Challenges and issues
5. Overview of tutorial
6. Unit summary
7. References

Background

- NLP has a 40+ year history of conducting human evaluation experiments to determine system quality.
- But very few established shared standards and methods for human evaluation.
- “There is so little control in individual tests and so much variation in method between tests that interpretations of the results of any one test or of their relationships with those of others must be uncertain.” (Sparck Jones, 1981, p. 245) – for human evaluation this is still true!

- Characterising evaluation experiments as D : M where D is the data and M the mechanism (system), Sparck Jones advocated the systematic testing of variations of D and M in a **uniform framework for system characterisation and evaluation**.
- D : M is now dominant paradigm, but with less variation in D than Sparck Jones envisaged.
- Still don’t have the “uniform framework for system characterisation and evaluation” advocated by Sparck Jones, strictly speaking not even for automatic evaluation.

Background

- Over recent years, issues and concerns have been raised many of which are due to this lack of a shared uniform framework.
- Will look at this in more detail, but some of the main issues and concerns are:
  - **Lack of standardisation** in what is being evaluated – does one evaluation of ‘Fluency’ assess the same thing as another? (Howcroft et al., 2020).
  - **Low levels of reproducibility** to the point where same main conclusions are often not supported by otherwise identical human evaluations (Belz et al., misc.).

  - **Poor practice in designing and executing experiments**, e.g. bugs, reporting errors, ad hoc interference in live experiments, etc. (Thomson et al., 2024).
  - **Loose application of experimental and statistical methods** and principles, e.g. not testing assumptions, unsuitable significance tests, no preregistration, over-reliance on post-hoc testing.

Background

- “our current understanding [...] is like that of sixteenth century herbalists: it embodies some observation and insight, but lacks detailed analysis and supporting theory.” (Sparck Jones, p. 3)
- There have periodically been calls for Sparck Jones’s “uniform framework for system characterisation and evaluation.”
- Increasing awareness in recent years that this is important, e.g. van der Lee et al. (2019), Howcroft et al. (2020) and Belz et al. (2020, 2021, 2022, 2023, 2024).
- Picking up proposals from the literature, this tutorial aims to:
    - Provide an overview of options and choices in human evaluation in NLP, with step-by-step guidance on how to put reliable experiments together.
    - Overall contribute to establishing standard best practice in human evaluation in NLP.

Overview

Unit 1: Introduction

1. Unit aims, learning outcomes, contents
2. What is human evaluation in NLP?
3. Background
4. Challenges and issues
5. Overview of tutorial
6. Unit summary
7. References

Diagnosing challenges and issues

- *Lack of standardisation* (van der Lee et al., 2019):
  - Wide variety of different quality criteria *names* in use, assessed in many different ways.
  - Many different tests of statistical significance and inter-annotator agreement applied.
  - Much variation in experimental design and reporting practices.
- *Unknown comparability* (Howcroft et al., 2020; Belz et al., 2020):
  - Extreme underreporting of details of evaluation studies → unclear if evaluations are comparable.
  - Even when quality criterion *names* are mapped to standard criteria, still 71 different quality criteria remain.

Criterion	Total	Criterion	Total
Fluency	40 (27%)	Readability	9 (6%)
Overall quality	29 (20%)	Appropriateness	7 (5%)
Informativeness	15 (10%)	Meaning preservation	6 (4%)
Relevance	15 (10%)	Clarity	5 (3%)
Grammaticality	14 (10%)	Non-reduncancy	4 (3%)
Naturalness	12 (8%)	Sentiment	4 (3%)
Coherence	10 (7%)	Consistency	4 (3%)
Accuracy	10 (7%)	Answerability	4 (3%)
Correctness	9 (6%)	Other criteria	124 (48%)*

Criterion Paraphrase	Count
usefulness for task/information need	39
grammaticality	39
quality of outputs	35
understandability	30
correctness of outputs relative to input (content)	29
goodness of outputs relative to input (content)	27
clarity	17
fluency	17
goodness of outputs in their own right	14
readability	14
information content of outputs	14
…	

Diagnosing challenges and issues

- *Low reproducibility* (Belz et al., 2021; Belz et al., 2023):
  - Systematic review of mostly metric evaluations assessed 549 individual score pairs, finding repeat experiments tend to produce worse scores.
  - For human evaluations, consistently poor reproducibility found for different quantitative and qualitative measures of reproducibility.

- *Poor experimental standards* (Thomson et al., 2024)

  In systematically selected set of 6 experiments, found flaws in all, some multiple times:
  - Response collection flaw
  - Inappropriate exclusion of responses: 3 cases
  - Reporting error: 4 cases
  - Coding error: 3 cases
  - Ethical flaw (here, failure to anonymise data)

Addressing challenges and issues

Addressing lack of standardisation and unknown comparability

- Recommendations for human evaluation best practice – van der Lee et al., 2019.
- QCET Taxonomy of standardised QCs (Belz et al., 2024).
- Human Evaluation Data Sheet (Shimorina & Belz, 2022) facilitates standard system and test characterisation (Sparck Jones).

Addressing poor experimental standards

- Recommendations to address flaws found in existing human evaluation experiments (Thomson et al., 2024).
- Automate experimental process as much as possible (Thomson & Belz, 2024).

Addressing challenges and issues

Addressing low reproducibility (Belz et al., 2021; Belz et al., 2023):

- Report *full* details in *standardised* from, e.g. HEDS (Shimorina & Belz, 2022).
- Use quantified measures to assess degree of reproducibility in comparable form, e.g. QRA++ (Belz & Thomson, 2024).
- Use techniques that have been shown to have better reproducibility.
- If feasible, test for reproducibility before running experiments (see Unit 7, Execution).

Over the course of the tutorial, will build on above recommendations, providing comprehensive details and guidance, in one place.

Overview

Unit 1: Introduction

1. Unit aims, learning outcomes, contents
2. What is human evaluation in NLP?
3. Background
4. Challenges and issues
5. Overview of tutorial
6. Unit summary
7. References

Overview of tutorial

- **Unit 1: Introduction**
- **Unit 2: Development and Components of Human Evaluations**
    - Standard terminology and definitions
    - Component processes of a (ready-to-run) human evaluation
    - Phases in creating and running a human evaluation
    - Example human evaluation in terms of the standard components
- **Unit 3: Quality Criteria and Evaluation Modes**
    - Research question(s) and hypotheses
    - Quality criteria and evaluation modes
    - Connection with formulating the research question
    - Practical task

Overview of tutorial

- **Unit 4: Experiment Design**
  - Design decisions and experiment properties
  - The 11 steps in designing a human evaluation experiment
  - Completing human evaluation data sheet (HEDS), with demo
- **Unit 5: Analysis**
  - Hypothesis testing
  - Pre-registration and statistical power analysis
  - Multiple hypothesis testing
  - Annotator reliability and representativeness
  - Post-hoc tests
  - Practical coding session

Overview of tutorial

- **Unit 6: Experiment Implementation**
  - From Design to Implementation
  - Implementing two remaining contributory component processes:
    - Sampling system outputs
    - Recruitment of evaluators
  - Implementing the five core component processes:
    - Normalising system outputs
    - Generation of evaluation interfaces
    - Response collection
    - Normalising raw responses
    - Analysis
  - Pipelining and documentation, updating the design
  - Good coding practices
  - Practical exercise

Overview of tutorial

- **Unit 7: Experiment Execution**
  - From implementation to execution
  - Pre-final execution
  - Preregistration and final execution
  - Ethics and fair treatment of participants
  - Post-final execution for reproducibility testing
- **Unit 8: Practical Session**
  - Exercises on response collection and normalisation
  - Exercises on results analysis

A note on scope

- Tutorial focuses on human evaluation of NLP system quality in a research context.
- We use the term *system outputs* for short, but intend it to be understood to also stand for sequences of outputs (as in dialogue), and other forms of user-system interaction records (as e.g. in extrinsic evaluation).
- Intended tutorial coverage is all types of evaluations of all types of systems encountered in NLP under this heading.
- Not included are:
    - Evaluation of efficiency, compute or memory requirements
    - Human evaluations in a product development context in industry
    - Full ethical assessment of evaluation experiments
    - Full assessment of systems in terms of social, ethical or environmental impact
    - Auditing of data or systems from legal or responsible AI perspective

Overview

Unit 1: Introduction

1. Unit aims, learning outcomes, contents
2. What is human evaluation in NLP?
3. Background
4. Challenges and issues
5. Overview of tutorial
6. Unit summary
7. References

Unit summary

- Human evaluation in NLP consists of experiments where participants are asked to rate or interact with NLP systems, and measures of quality are collected.
- A typical evaluation scenario in NLP involves collecting a representative sample of outputs from a set of comparable systems, and asking participants to rate them according to given quality criteria.
- What constitutes good system quality depends on general output properties such as their fluency and grammaticality, as well as application-specific factors such as what is appropriate for given situations or purposes.
- Evaluation methods in use in NLP vary to the point that interpretations of the results of any one evaluation or of their relationships with others are uncertain (Sparck Jones).
- Goal is a uniform framework for system characterisation and evaluation (Sparck Jones), but the field remains far from this goal.
- The main challenges and issues that have been identified are: lack of standardisation, unclear comparability of evaluations, low reproducibility, and poor experimental standards.

Unit summary

- This tutorial builds on existing work to address these challenges and issues, aiming to provide a complete, one-stop guide to creating and running scientifically rigorous human evaluations of NLP system quality.
- The scope of the tutorial is all types of human evaluations of the quality of all types of NLP systems in a research context.
- This scope excludes assessments of other aspects such as computational efficiency, social and environmental impact, and AI responsibility.

Overview

Unit 1: Introduction

1. Unit aims, learning outcomes, contents
2. What is human evaluation in NLP?
3. Background
4. Challenges and issues
5. Current research in human evaluation of NLP systems
6. Overview of tutorial
7. Unit summary
8. References

References

Essential:

Human evaluation of automatically generated text: Current trends and best practice guidelines.
Chris van der Lee, Albert Gatt, Emiel van Miltenburg, Emiel Krahmer. Computer Speech & Language, Volume 67, 2021.

Twenty Years of Confusion in Human Evaluation: NLG needs evaluation sheets and standardised definitions.
D Howcroft, A Belz, D Gkatzia, S Hasan, S Mahamood, S Mille, M Clinciu et al. International Natural Language Generation Conference 2020 (INLG'20).

Non-repeatable experiments and non-reproducible results: The reproducibility crisis in human evaluation in NLP.
A Belz, C Thomson, E Reiter, S Mille. Findings of ACL 2023, 3676-3687, 2023.

Common Flaws in Running Human Evaluation Experiments in NLP
C Thomson, E Reiter, A Belz. Computational Linguistics, 1-11, 2024.

References

Further reading:

Disentangling the Properties of Human Evaluation Methods: A Classification System to Support Comparability, Meta-Evaluation and Reproducibility Testing.
A Belz, S Mille, D Howcroft. International Natural Language Generation Conference 2020 (INLG'20).

A Systematic Review of Reproducibility Research in Natural Language Processing.
A Belz, S Agarwal, A Shimorina, E Reiter. *EACL'21*.

Information retrieval experiment.
Karen Sparck Jones, Butterworth-Heinemann, 1981.

Best practices for the human evaluation of automatically generated text.
Chris van der Lee, Albert Gatt, Emiel van Miltenburg, Sander Wubben, and Emiel Krahmer.
International Natural Language Generation Conference 2019 (INLG'19).