Flamingos and Hedgehogs in the Croquet-Ground: Teaching Evaluation of NLP Systems for Undergraduate Students

Brielen Madureira
Computational Linguistics
Department of Linguistics
University of Potsdam
madureiralasota@uni-potsdam.de

Abstract

This report describes the course *Evaluation of NLP Systems*, taught for Computational Linguistics undergraduate students during the winter semester 20/21 at the University of Potsdam, Germany. It was a discussion-based seminar that covered different aspects of evaluation in NLP, namely paradigms, common procedures, data annotation, metrics and measurements, statistical significance testing, best practices and common approaches in specific NLP tasks and applications.

1 Motivation

"Alice soon came to the conclusion that it was a very difficult game indeed." [1]

When the Queen of Hearts invited Alice to her croquet-ground, Alice had no idea how to play that strange game with flamingos and hedgehogs. NLP newcomers may be as puzzled as her when they enter the Wonderland of NLP and encounter a myriad of strange new concepts: Baseline, F1 score, glass box, ablation, diagnostic, extrinsic and intrinsic, performance, annotation, metrics, human-based, test suite, shared task...

Although experienced researchers and practitioners may easily relate them to the evaluation of NLP models and systems, for newcomers like undergraduate students it is not simply a matter of looking up their definition. It is necessary to show them the big picture of what and how we play in the croquet-ground of evaluation in NLP.

The NLP community clearly cares for doing proper evaluation. From earlier works like the book by Karen Spärck Jones and Julia R. Galliers (1995) to the winner of ACL 2020 best paper award (Ribeiro et al., 2020) and recent dedicated workshops, *e.g.* Eger et al. (2020), the formulation of evaluation methodologies has been a prominent topic in the field.

Despite its importance, evaluation is usually covered very briefly in NLP courses due to a tight schedule. Teachers barely have time to discuss dataset splits, simple metrics like accuracy, precision, recall and F1 Score, and some techniques like cross validation. As a result, students end up learning about evaluation on-the-fly as they begin their careers in NLP. The lack of structured knowledge may cause them to be unacquainted with the multi-faceted metrics and procedures, which can render them partially unable to evaluate models critically and responsibly. The leap from that one lecture to what is expected in good NLP papers and software should not be underestimated.

The course *Evaluation of NLP Systems*, which I taught for undergraduate Computational Linguistics students in the winter semester of 20/21 at the University of Potsdam, Germany, was a reading and discussion-based learning approach with three main goals: i) helping participants become aware of the importance of evaluation in NLP; ii) discussing different evaluation methods, metrics and techniques; and iii) showing how evaluation is being done for different NLP tasks.

The following sections provide an overview of the course content and structure. With some adaptation, this course can also be suitable for more advanced students.

[1] Alice in Wonderland by Lewis Carroll, public domain. Illustration by John Tenniel, public domain, via Wikimedia Commons.

Proceedings of the Fifth Workshop on Teaching NLP, pages 87–91

Topic	Content
Paradigms	Kinds of evaluation and main steps, *e.g.* intrinsic and extrinsic, manual and automatic, black box and glass box.
Common Procedures	Overview about the use of measurements, baselines, dataset splits, cross validation, error analysis, ablation, human evaluation and comparisons.
Annotation	How to annotate linguistic data, evaluate the annotation and how the annotation scheme can affect the evaluation of a system's performance.
Metrics and Measurements	Outline of the different metrics commonly used in NLP, what they aim to quantify and how to interpret them.
Statistical Significance Testing	Hypothesis testing for comparing the performance of two systems in the same dataset.
Best Practices	The linguistic aspect of NLP, reproducibility and the social impact of NLP.
NLP Case Studies	Group presentations about specific approaches in four NLP tasks/applications (machine translation, natural language generation, dialogue and speech synthesis) and related themes (the history of evaluation, shared tasks, ethics and ACL's code of conduct and replication crisis).

Table 1: Overview of the course content.

2 Course Content and Format

Table 1 presents an overview of the topics discussed in the course. Details about the weekly reading lists are available at the course's website.[2]

The course happened 100% online due to the pandemic. It was divided into two parts. In the first half of the semester, students learned about the evaluation methods used in general in NLP and, to some extent, machine learning. After each meeting, I posted a pre-recorded short lecture, slides and a reading list about the next week's content. The participants had thus one week to work through the material anytime before the next meeting slot. I provided diverse sources like papers, blogposts, tutorials, slides and videos.

I started the online meetings with a wrap-up and feedback about the previous week's content. Then, I randomly split them into groups of 3 or 4 participants in breakout sessions so that they could discuss a worksheet together for about 45 minutes. I encouraged them to use this occasion to profit from the interaction and brainstorming with their peers and exchange arguments and thoughts. After the meeting, they had one week to write down their solutions individually and submit it.

In the second half of the semester, they divided into 4 groups to analyze how evaluation is being done in specific NLP tasks. For larger groups, other NLP tasks can be added. They prepared group presentations and discussion topics according to general guidelines and an initial bibliography that they could expand. Students provided anonymous feedback about each other's presentations for me and I then shared it with the presenters, to have the chance to filter abusive or offensive comments.

The last lecture was a tutorial about useful metrics available in *scikit-learn* and *nltk* Python libraries using Jupyter Notebook (Kluyver et al., 2016).

Finally, they had six weeks to work on a final project. Students could select one of the following three options: i) a critical essay on the development and current state of evaluation in NLP, discussing the positive and negative aspects and where to go from here; ii) a hands-on detailed evaluation of an NLP system of their choice, which could be,

[2] https://briemadu.github.io/evalNLP/schedule

for example, an algorithm they implemented for another course; or iii) a summary of the course in the format of a small newspaper.

3 Participants

Seventeen bachelor students of Computational Linguistics attended the course. At the University of Potsdam, this seminar falls into the category of a module called *Methods of Computational Linguistics*, which is intended for students in the 5th semester of their bachelor course. Still, one student in the 3rd and many students in higher semesters also took part.

By the 5th semester, students are expected to have completed introductory courses on linguistics (phonetic and phonology, syntax, morphology, semantics and psycho- and neurolinguistics), computational linguistics techniques, computer science and programming (finite state automata, advanced Python and other courses of their choice), introduction to statistics and empirical methods and foundations of mathematics and logic, as well as varying seminars related to computational linguistics.

Although there were no formal requirements for taking this course, students should preferably be familiar some common tasks and practices in NLP and the basics of statistics.

4 Outcomes

I believe this course successfully introduced students to several fundamental principles of evaluation in NLP. The quality of their submissions, especially the final project, was, in general, very high. By knowing how to properly manage flamingos and hedgehogs, they will hopefully be spared the sentence "*off with their head!*" as they continue their careers in NLP. The game is not very difficult when one learns the rules.

Students gave very positive feedback at the end of the semester about the content, the literature and the format. They particularly enjoyed the opportunity to discuss with each other, saying it was good to exchange what they recalled from the reading. They also stated that what they learned contributed to their understanding in other courses and improved their ability to document and evaluate models they implement. The course was also useful for them to start reading more scientific literature.

In terms of improvements, they mentioned that the weekly workload could be reduced. They also reported that the reading for the week when we covered statistical significance testing was too advanced. Still, they could do the worksheet since it did not dive deep into the theory.

The syllabus, slides and suggested readings are available on the course's website.[3] The references here list the papers and books used to put together the course and has no ambition of being exhaustive. In case this course is replicated, the references should be updated with the most recent papers. I can share the worksheets and guidelines for the group presentation and the project upon request. Feedback from readers is very welcome.

Acknowledgments

In this course, I was inspired and used material available online by many people, to whom I am thankful. I also thank the students who were very engaged during the semester and made it a rewarding experience for me. Moreover, I am grateful for the anonymous reviewers for their detailed and encouraging feedback.

References

Valerie Barr and Judith L. Klavans. 2001. Verification and validation of language processing systems: Is it evaluation? In *Proceedings of the ACL 2001 Workshop on Evaluation Methodologies for Language and Dialogue Systems*.

Yonatan Belinkov and James Glass. 2019. Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 7:49–72.

Anja Belz. 2009. That's nice… what can you do with it? *Computational Linguistics*, 35(1):111–118.

Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Taylor Berg-Kirkpatrick, David Burkett, and Dan Klein. 2012. An empirical investigation of statistical significance in NLP. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 995–1005, Jeju Island, Korea. Association for Computational Linguistics.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2007.

[3] https://briemadu.github.io/evalNLP/

(meta-) evaluation of machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 136–158, Prague, Czech Republic. Association for Computational Linguistics.

Asli Celikyilmaz, Elizabeth Clark, and Jianfeng Gao. 2020. Evaluation of text generation: A survey. *arXiv preprint arXiv:2006.14799*.

Eirini Chatzikoumi. 2020. How to evaluate machine translation: A review of automated and human metrics. *Natural Language Engineering*, 26(2):137–161.

Jan Deriu, Alvaro Rodrigo, Arantxa Otegi, Guillermo Echegoyen, Sophie Rosset, Eneko Agirre, and Mark Cieliebak. 2021. Survey on evaluation methods for dialogue systems. *Artificial Intelligence Review*, 54(1):755–810.

Rotem Dror, Gili Baumer, Marina Bogomolov, and Roi Reichart. 2017. Replicability analysis for natural language processing: Testing significance with multiple datasets. *Transactions of the Association for Computational Linguistics*, 5:471–486.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1383–1392, Melbourne, Australia. Association for Computational Linguistics.

Rotem Dror, Lotem Peled-Cohen, Segev Shlomov, and Roi Reichart. 2020. Statistical significance testing for natural language processing. *Synthesis Lectures on Human Language Technologies*, 13(2):1–116.

Steffen Eger, Yang Gao, Maxime Peyrard, Wei Zhao, and Eduard Hovy, editors. 2020. *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*. Association for Computational Linguistics, Online.

Karën Fort, Gilles Adda, and K. Bretonnel Cohen. 2011. Last words: Amazon Mechanical Turk: Gold mine or coal mine? *Computational Linguistics*, 37(2):413–420.

J.R. Galliers, K.S. Jones, and University of Cambridge. Computer Laboratory. 1993. *Evaluating Natural Language Processing Systems*. Computer Laboratory Cambridge: Technical report. University of Cambridge, Computer Laboratory.

Albert Gatt and Emiel Krahmer. 2018. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, 61:65–170.

Kyle Gorman and Steven Bedrick. 2019. We need to talk about standard splits. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2786–2791, Florence, Italy. Association for Computational Linguistics.

Lynette Hirschman and Henry S. Thompson. 1997. *Overview of Evaluation in Speech and Natural Language Processing*, page 409–414. Cambridge University Press, USA.

Dirk Hovy and Shannon L. Spruit. 2016. The social impact of natural language processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 591–598, Berlin, Germany. Association for Computational Linguistics.

David M. Howcroft, Anya Belz, Miruna-Adriana Clinciu, Dimitra Gkatzia, Sadid A. Hasan, Saad Mahamood, Simon Mille, Emiel van Miltenburg, Sashank Santhanam, and Verena Rieser. 2020. Twenty years of confusion in human evaluation: NLG needs evaluation sheets and standardised definitions. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 169–182, Dublin, Ireland. Association for Computational Linguistics.

Karen Sparck Jones and Julia R Galliers. 1995. *Evaluating natural language processing systems: An analysis and review*, volume 1083 of *Lecture Notes in Artificial Intelligence*. Springer-Verlag Berlin Heidelberg.

Margaret King. 1996. Evaluating natural language processing systems. *Communications of the ACM*, 39(1):73–79.

Thomas Kluyver, Benjamin Ragan-Kelley, Fernando Pérez, Brian Granger, Matthias Bussonnier, Jonathan Frederic, Kyle Kelley, Jessica Hamrick, Jason Grout, Sylvain Corlay, Paul Ivanov, Damián Avila, Safia Abdalla, and Carol Willing. 2016. Jupyter notebooks – a publishing format for reproducible computational workflows. In *Positioning and Power in Academic Publishing: Players, Agents and Agendas*, pages 87 – 90. IOS Press.

Zachary C. Lipton and Jacob Steinhardt. 2019. Troubling trends in machine learning scholarship: Some ml papers suffer from flaws that could mislead the public and stymie future research. *Queue*, 17(1):45–77.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132, Austin, Texas. Association for Computational Linguistics.

Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. 2017. Why we need new evaluation metrics for NLG. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2241–2252, Copenhagen, Denmark. Association for Computational Linguistics.

Patrick Paroubek, Stéphane Chaudiron, and Lynette Hirschman. 2007. Principles of evaluation in natural language processing. *Traitement Automatique des Langues*, 48(1):7–31.

Carla Parra Escartín, Wessel Reijers, Teresa Lynn, Joss Moorkens, Andy Way, and Chao-Hong Liu. 2017. Ethical considerations in NLP shared tasks. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 66–73, Valencia, Spain. Association for Computational Linguistics.

Ehud Reiter and Anja Belz. 2009. An investigation into the validity of some metrics for automatically evaluating natural language generation systems. *Computational Linguistics*, 35(4):529–558.

Philip Resnik and Jimmy Lin. 2010. Evaluation of NLP systems. *The handbook of computational linguistics and natural language processing. Chapter 11.*, 57.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.

Stefan Riezler and John T. Maxwell. 2005. On some pitfalls in automatic evaluation and significance testing for MT. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 57–64, Ann Arbor, Michigan. Association for Computational Linguistics.

Nathan Schneider. 2015. What I've learned about annotating informal text (and why you shouldn't take my word for it). In *Proceedings of The 9th Linguistic Annotation Workshop*, pages 152–157, Denver, Colorado, USA. Association for Computational Linguistics.

Noah A Smith. 2011. Linguistic structure prediction. *Synthesis lectures on human language technologies*, 4(2):1–274.

Anders Søgaard, Anders Johannsen, Barbara Plank, Dirk Hovy, and Hector Martínez Alonso. 2014. What's in a p-value in NLP? In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pages 1–10, Ann Arbor, Michigan. Association for Computational Linguistics.

Karen Sparck Jones. 1994. Towards better NLP system evaluation. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*.

Chris van der Lee, Albert Gatt, Emiel van Miltenburg, Sander Wubben, and Emiel Krahmer. 2019. Best practices for the human evaluation of automatically generated text. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 355–368, Tokyo, Japan. Association for Computational Linguistics.

Petra Wagner, Jonas Beskow, Simon Betz, Jens Edlund, Joakim Gustafson, Gustav Eje Henter, Sébastien Le Maguer, Zofia Malisz, Éva Székely, Christina Tånnander, et al. 2019. Speech synthesis evaluation—state-of-the-art assessment and suggestion for a novel research program. In *Proceedings of the 10th Speech Synthesis Workshop (SSW10)*.