A Tutorial on Evaluation Metrics used in Natural Language Generation

Mitesh M. Khapra and **Ananya B. Sai**
Robert-Bosch Centre for Data Science & Artificial Intelligence
Indian Institute of Technology Madras
India
{miteshk,ananya}@cse.iitm.ac.in

Abstract

There has been a massive surge of Natural Language Generation (NLG) models in the recent years, accelerated by deep learning and the availability of large-scale datasets. With such rapid progress, it is vital to assess the extent of scientific progress made and identify the areas/components that need improvement. To accomplish this in an automatic and reliable manner, the NLP community has actively pursued the development of automatic evaluation metrics. Especially in the last few years, there has been an increasing focus on evaluation metrics, with several criticisms of existing metrics and proposals for several new metrics. This tutorial presents the evolution of automatic evaluation metrics to their current state along with the emerging trends in this field by specifically addressing the following questions: (i) What makes NLG evaluation challenging? (ii) Why do we need *automatic* evaluation metrics? (iii) What are the existing automatic evaluation metrics and how can they be organised in a coherent taxonomy? (iv) What are the criticisms and shortcomings of existing metrics? (v) What are the possible future directions of research?

1 Tutorial Content Description

Natural Language Generation (NLG) encompasses various tasks that require an automatic generation of human-understandable text such as Machine Translation, Abstractive Summarization, Question Answering, Data-to-text Generation, Dialogue Systems, etc. Each of these tasks has several use-cases with numerous models proposed over the years. The successful application of machine learning and deep learning techniques has transformed the mainstream models for NLG from rule-based systems to data-driven, end-to-end trainable systems. The easier availability of datasets and access to powerful computing resources has led to the wide-spread adoption of these techniques and rapid developments in the field. To track the developments and understand the scientific progress made, these NLG systems need to be evaluated carefully. The ideal way to do so would be to employ expert human evaluators. However, this option would be very time-consuming and expensive, and is thus infeasible. Hence the community has settled for automatic evaluation metrics to track scientific progress in this field.

Automatic Evaluation metrics such as BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), ROUGE (Lin, 2004) have been around for several years and are still predominantly used. They have also been readily adopted for newer tasks in NLG such as Question Generation, Image Captioning, etc, due to the lack of any other relevant metrics. However, there has been heavy criticism for such an adoption of metrics across tasks, corroborated by their poor correlations with human judgements (Liu et al., 2016; Nema and Khapra, 2018; Dhingra et al., 2019). Several new metrics are being proposed to address the shortcomings of the existing ones (Sai et al., 2020b). The emerging metrics also explore the idea of using the context provided for the task (such as a document, image, passage, or tabular data, and so on), unlike BLEU, METEOR, ROUGE, etc. This has lead to the development of 'context-dependent metrics' alongside the 'context-free metrics'.

Both the context-free and context-dependent metrics can be categorized based on their underlying technique into trained metrics and untrained (i.e., rule-based/ heuristic-based) metrics. Untrained metrics can be further classified depending on whether they are word-based (Papineni et al., 2002; Banerjee and Lavie, 2005; Lin, 2004; Snover et al., 2006; Druck and Pang, 2012; Dhingra et al., 2019), character-based (Popovic, 2015; Wang et al., 2016), or embedding-based (Rus and Lintean, 2012; Forgues et al., 2014; Kusner et al., 2015; Mathur et al., 2019; Zhang et al., 2019). Similarly, trained metrics are sub-categorized depending on

Proceedings of NAACL-HLT 2021: Tutorials, pages 15–19

whether they need input features (such as precision, recall, number of words in a sentence, etc,) (Stanojevic and Sima'an, 2014; Ma et al., 2017; Nema and Khapra, 2018) or whether they extract the features from the input sentences in an end-to-end manner (Lowe et al., 2017; Tao et al., 2018; Cui et al., 2018; Shimanaka et al., 2018; Wieting et al., 2019; Sellam et al., 2020; Sai et al., 2020a). In this tutorial, we provide an overview of these different techniques that have been used to formulate automatic evaluation metrics. We also discuss the studies that analyze/inspect these metrics and report their shortcomings. The major criticisms on the metrics include the uninterpretability of the scores (Zhang et al., 2004; Callison-Burch et al., 2006), bias towards specific models (Dusek et al., 2020) or scores (Sai et al., 2019), and their inability to capture all the nuances in a task (Ananthakrishnan et al., 2006). We conclude by presenting the possible next directions of research in automatic evaluation metrics.

1.1 Relevance to computational linguistics community

There is a necessity to compare the myriad of models being proposed for various NLG tasks and scrutinize the progress carefully. Towards this objective, the topic of evaluation metrics has been highly relevant to the linguistics community, in general, and to researchers working on various tasks in NLG, in particular. The number of research papers that critically examine the existing metrics and/or propose new metrics has been rapidly increasing. For example, at least 40 new metrics have been proposed since 2014 for various NLG tasks. We thus believe that the topic of automatic evaluation metrics is garnering more interest in the recent years. This tutorial aims to bring new and existing researchers up-to-speed on the developments related to this topic.

2 Type of the Tutorial

Cutting-edge: This tutorial will follow the growth of automatic evaluation metrics over the years, starting with the initial metrics that are still popularly used today, and building up to the more recent metrics. Substantial emphasis will be given to the recent trends and emerging directions of research on this topic. To the best of our knowledge such a tutorial on evaluation metrics has not been conducted so far in any of the ACL/EACL/IJCNLP/EMNLP/NAACL venues.

3 Tutorial Structure and Schedule Outline

We plan a 3 hour tutorial based on the following content and associated time estimates.

- Introduction (20 min)
  - NLG (A brief history)
  - Have we made progress?
  - Quantifying progress
    - Human/Manual Evaluation
    - Automatic Evaluation
  - Tutorial Roadmap
- Challenges of Automatic Evaluation of NLG tasks (20 min)
  - Breakdown of evaluation criteria for different tasks
    - Machine Translation
    - Abstractive Summarization
    - Question Answering
    - Question Generation
    - Data-to-Text Generation
    - Dialogue Generation
    - Image Captioning
  - Summary of the Challenges
- Taxonomy of Automatic Evaluation Metrics in use (10 min)
  - Context-free v/s Context-dependent metrics
  - Trained metrics v/s Untrained (/heuristic-based) metrics
  - Task-specific v/s Task-agnostic metrics
- Context-free metrics (30 min)
  - Untrained metrics
    - Word or character based metrics
    - Embedding based metrics
  - Trained metrics
    - Feature-based metrics
    - End-to-end trained metrics
- Context-dependent metrics (30 min)
  - Untrained metrics
  - Trained metrics

- Shortcomings identified in existing metrics (40 min)
  - Poor correlations
  - Uninterpretability of scores
  - Bias in the metrics
  - Poor adaptability across tasks
  - Inability to capture all nuances in a task
- Conclusions and future research directions (10 min)

4 Prerequisites

We aim to present the tutorial in a self-contained manner, accommodating audience with various backgrounds. However, it would be helpful to have basic knowledge about Natural Language Processing, Machine Learning, and Deep Learning methods (such as Word embeddings, Recurrent Neural Networks, Sequence-to-sequence models, and Transformers).

5 Presenters

Mitesh M. Khapra, Assistant Professor, Indian Institute of Technology Madras
Email: miteshk@cse.iitm.ac.in
Site: http://www.cse.iitm.ac.in/~miteshk/
Mitesh M. Khapra is an Assistant Professor in the Department of Computer Science and Engineering at IIT Madras and is affiliated with the Robert Bosch Centre for Data Science and AI. He co-founded One Fourth Labs, with a mission to design and deliver affordable hands-on courses on AI and related topics. He is also a co-founder of AI4Bharat, a voluntary community with an aim to provide AI-based solutions to India-specific problems. His research interests span the areas of Deep Learning, Multimodal Multilingual Processing, Natural Language Generation, Dialog systems, Question Answering and Indic Language Processing. He has publications in several top conferences and journals including TACL, ACL, NeurIPS, ICLR, EMNLP, EACL, AAAI, etc. He has also served as Area Chair or Senior PC member in top conferences such as ICLR and AAAI. Prior to IIT Madras, he worked as a Researcher at IBM Research India for four and half years. While at IBM, he worked on several interesting problems in the areas of Statistical Machine Translation, Cross Language Learning, Multimodal Learning, Argument Mining and Deep Learning. Prior to IBM, he completed his PhD and M.Tech from IIT Bombay in Jan 2012 and July 2008 respectively. His PhD thesis dealt with the important problem of reusing resources for multilingual computation. During his PhD he was a recipient of the IBM PhD Fellowship (2011) and the Microsoft Rising Star Award (2011). He is also a recipient of the Google Faculty Research Award (2018), the IITM Young Faculty Recognition Award (2019), and the Prof. B. Yegnanarayana Award for Excellence in Research and Teaching (2020). He has previously presented tutorials at NAACL 2016 on "Multilingual Multimodal Language Processing Using Neural Networks" and "Statistical Machine Translation between Related Languages".

Ananya B. Sai, PhD student, Indian Institute of Technology Madras
Email: ananya@cse.iitm.ac.in
Site: https://ananyasaib.github.io/
Ananya Sai is currently a PhD student in the Department of Computer Science and Engineering at IIT Madras working with Dr. Mitesh M. Khapra. Her research interests include Natural Language Processing, Deep Learning, Adversarial Attacks, and Dialog Systems. Some of her recent research works are related to adversarial attacks on trained evaluation models. These include whitebox attacks and synthetic or human crafted adversarial modifications of the input sentences to fool the models. She has co-created a multi-reference dialogue dataset and has explored the benefits of task-specific pretraining for evaluating dialogue systems. She is a recipient of Google PhD Fellowship (2019) and the Prime Minister Fellowship for Doctoral Research (2020). She has published papers in TACL, AAAI, and IJCAI.

References

Ananthakrishnan, Pushpak Bhattacharyya, Murugesan Sasikumar, and Ritesh M. Shah. 2006. Some issues in automatic evaluation of english-hindi mt : More blues for BLEU. In *Proceeding of 5th International Conference on Natural Language Processing (ICON-07). Hyderabad, India.*

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Ar-

bor, Michigan. Association for Computational Linguistics.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the role of Bleu in machine translation research. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, Trento, Italy. Association for Computational Linguistics.

Yin Cui, Guandao Yang, Andreas Veit, Xun Huang, and Serge J. Belongie. 2018. Learning to evaluate image captioning. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 5804–5812. IEEE Computer Society.

Bhuwan Dhingra, Manaal Faruqui, Ankur P. Parikh, Ming-Wei Chang, Dipanjan Das, and William W. Cohen. 2019. Handling divergent reference texts when evaluating table-to-text generation. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 4884–4895. Association for Computational Linguistics.

Gregory Druck and Bo Pang. 2012. Spice it up? mining refinements to online instructions from user generated content. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 545–553, Jeju Island, Korea. Association for Computational Linguistics.

Ondrej Dusek, Jekaterina Novikova, and Verena Rieser. 2020. Evaluating the state-of-the-art of end-to-end natural language generation: The E2E NLG challenge. *Comput. Speech Lang.*, 59:123–156.

Gabriel Forgues, Joelle Pineau, Jean-Marie Larchevêque, and Réal Tremblay. 2014. Bootstrapping dialog systems with word embeddings. In *Neurips, modern machine learning and natural language processing workshop*, volume 2.

Matt J. Kusner, Yu Sun, Nicholas I. Kolkin, and Kilian Q. Weinberger. 2015. From word embeddings to document distances. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 957–966. JMLR.org.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Michael 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, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 2122–2132. The Association for Computational Linguistics.

Ryan Lowe, Michael Noseworthy, Iulian Vlad Serban, Nicolas Angelard-Gontier, Yoshua Bengio, and Joelle Pineau. 2017. Towards an automatic turing test: Learning to evaluate dialogue responses. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 1116–1126. Association for Computational Linguistics.

Qingsong Ma, Yvette Graham, Shugen Wang, and Qun Liu. 2017. Blend: a novel combined MT metric based on direct assessment - CASICT-DCU submission to WMT17 metrics task. In *Proceedings of the Second Conference on Machine Translation, WMT 2017, Copenhagen, Denmark, September 7-8, 2017*, pages 598–603. Association for Computational Linguistics.

Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 2019. Putting evaluation in context: Contextual embeddings improve machine translation evaluation. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 2799–2808. Association for Computational Linguistics.

Preksha Nema and Mitesh M. Khapra. 2018. Towards a better metric for evaluating question generation systems. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 3950–3959. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Maja Popovic. 2015. chrf: character n-gram f-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation, WMT@EMNLP 2015, 17-18 September 2015, Lisbon, Portugal*, pages 392–395. The Association for Computer Linguistics.

Vasile Rus and Mihai C. Lintean. 2012. A comparison of greedy and optimal assessment of natural language student input using word-to-word similarity metrics. In *BEA@NAACL-HLT*.

Ananya B. Sai, Mithun Das Gupta, Mitesh M. Khapra, and Mukundhan Srinivasan. 2019. Re-evaluating ADEM: A deeper look at scoring dialogue responses. In *The Thirty-Third AAAI Conference on Artificial*

Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019, pages 6220–6227. AAAI Press.

Ananya B. Sai, Akash Kumar Mohankumar, Siddhartha Arora, and Mitesh M. Khapra. 2020a. Improving dialog evaluation with a multi-reference adversarial dataset and large scale pretraining. *Trans. Assoc. Comput. Linguistics*, 8:810–827.

Ananya B. Sai, Akash Kumar Mohankumar, and Mitesh M. Khapra. 2020b. A survey of evaluation metrics used for NLG systems. *CoRR*, abs/2008.12009.

Thibault Sellam, Dipanjan Das, and Ankur P. Parikh. 2020. BLEURT: learning robust metrics for text generation. *CoRR*, abs/2004.04696.

Hiroki Shimanaka, Tomoyuki Kajiwara, and Mamoru Komachi. 2018. RUSE: regressor using sentence embeddings for automatic machine translation evaluation. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers, WMT 2018, Belgium, Brussels, October 31 - November 1, 2018*, pages 751–758. Association for Computational Linguistics.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *In Proceedings of Association for Machine Translation in the Americas*, pages 223–231.

Milos Stanojevic and Khalil Sima'an. 2014. Fitting sentence level translation evaluation with many dense features. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 202–206. ACL.

Chongyang Tao, Lili Mou, Dongyan Zhao, and Rui Yan. 2018. RUBER: an unsupervised method for automatic evaluation of open-domain dialog systems. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 722–729. AAAI Press.

Weiyue Wang, Jan-Thorsten Peter, Hendrik Rosendahl, and Hermann Ney. 2016. Character: Translation edit rate on character level. In *Proceedings of the First Conference on Machine Translation, WMT 2016, colocated with ACL 2016, August 11-12, Berlin, Germany*, pages 505–510. The Association for Computer Linguistics.

John Wieting, Taylor Berg-Kirkpatrick, Kevin Gimpel, and Graham Neubig. 2019. Beyond BLEU: training neural machine translation with semantic similarity. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 4344–4355. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *ArXiv*, abs/1904.09675.

Ying Zhang, Stephan Vogel, and Alex Waibel. 2004. Interpreting BLEU/NIST scores: How much improvement do we need to have a better system? In *Proceedings of the Fourth International Conference on Language Resources and Evaluation, LREC 2004, May 26-28, 2004, Lisbon, Portugal*. European Language Resources Association.