Judge the Judges: A Large-Scale Evaluation Study of Neural Language Models for Online Review Generation

Cristina Garbacea (garbacea@umich.edu)[1], Samuel Carton[2], Shiyan Yan[2] and Qiaozhu Mei[1,2]
[1]EECS Department, [2]School of Information, University of Michigan, Ann Arbor, MI, USA

Background & Motivation

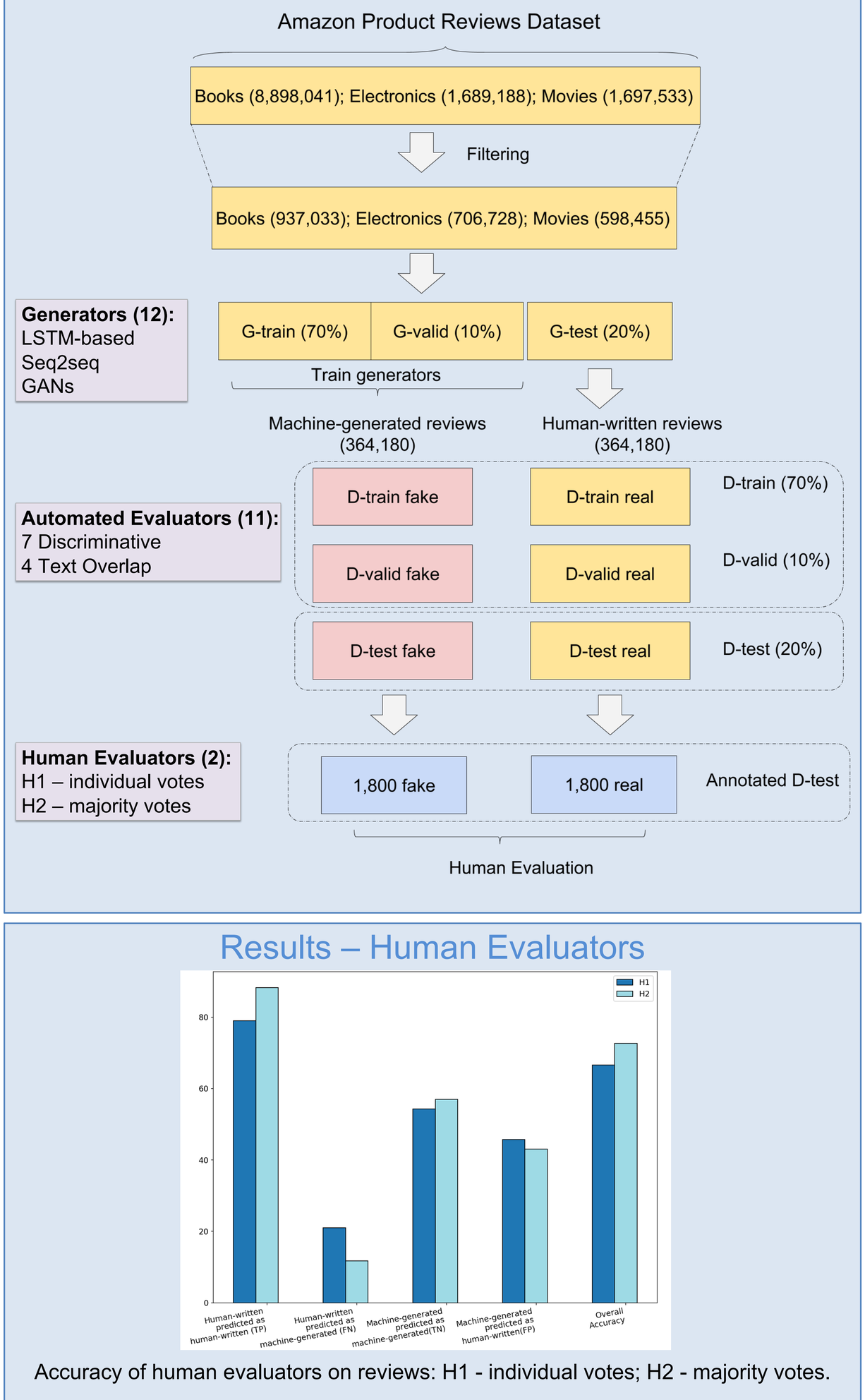
Evaluation for Natural Language Generation (NLG):
- no agreed objective criterion or clear model of text quality for comparing the "goodness" of generated texts
- the most natural way to evaluate the quality of a generator is to involve humans as judges (Turing test), however such approaches are hard to scale
- it is critical to find automated metrics to evaluate the quality of a generator independent of human judges or an exhaustive set of competing generators

Present Work

- We design a large-scale experiment to **evaluate the evaluators for NLG**.
- We compare three types of evaluators: *human evaluators*, *automated adversarial evaluators* trained to distinguish human-written from machine-generated texts, and *word overlap metrics*.

Experimental Setup

- **Task:** classify a product review as human-written or machine-generated
- **Dataset:** Amazon Product Reviews dataset (books, electronics, movies)

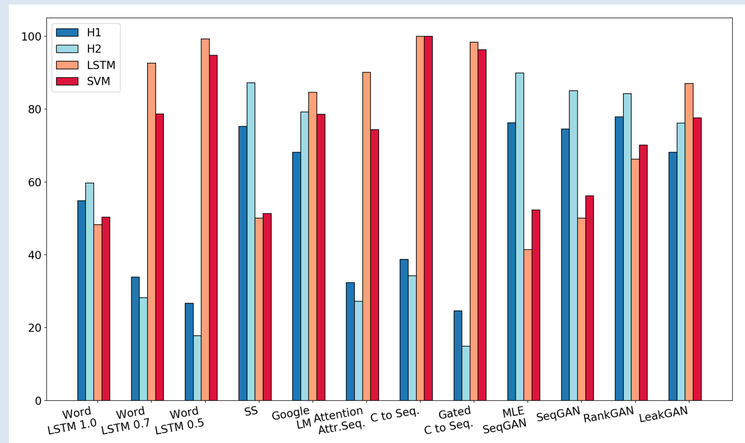


Results – Human Evaluators

Accuracy of human evaluators on reviews: H1 - individual votes; H2 - majority votes.

Human evaluators generally do better at correctly labeling human-written reviews as real (H1 - 78.96%, H2 - 88.31%), but are confused by machine-generated reviews in close to half of the cases (H1 - 54.26%, H2 - 56.95%).

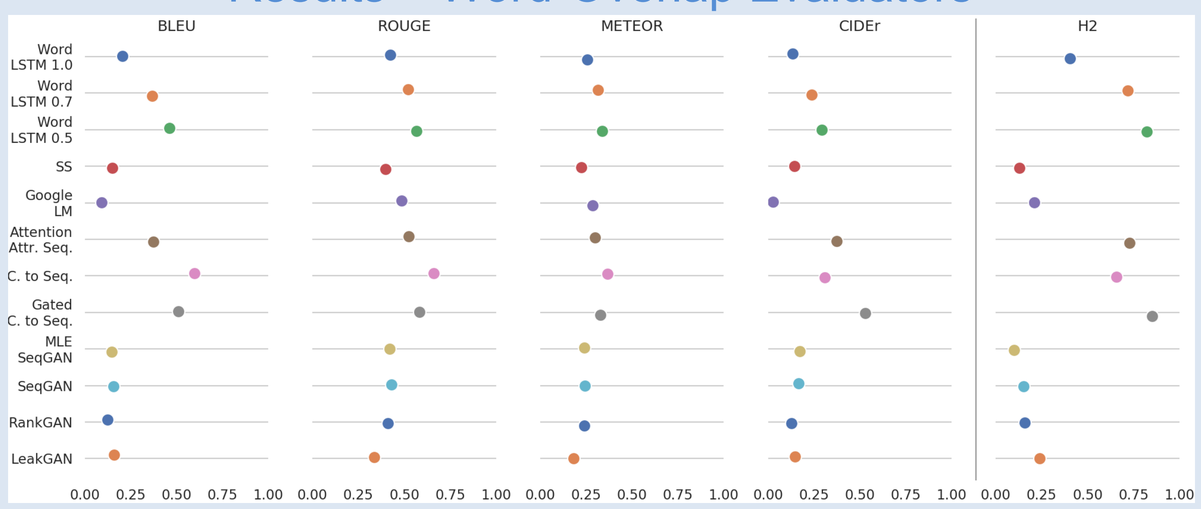
Results – Discriminative Evaluators



Accuracy of human (H1, H2) and discriminative evaluators (LSTM, SVM) on reviews. **The lower the accuracy, the better the generator.**

Discriminative evaluators can distinguish a single machine-generated review from human-written better than humans, GAN generators are ranked highest. **Human evaluators** are not fooled by GAN-generated reviews.

Results – Word-Overlap Evaluators

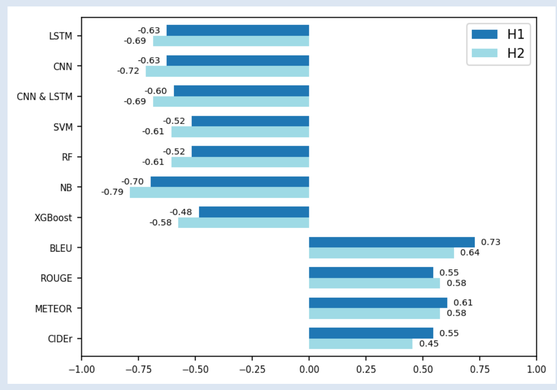


Text-Overlap Evaluators scores for individual generators. **The higher the better.**

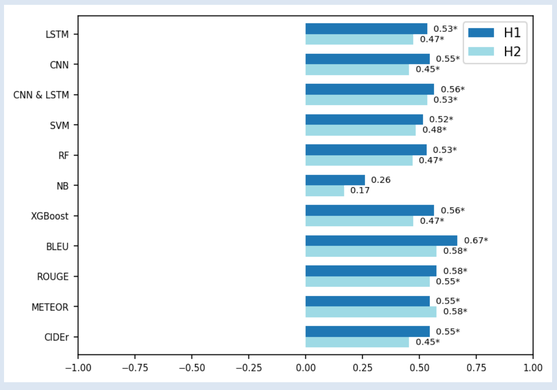
Different word-overlap evaluators rank generators in similar order: seq2seq generators are ranked highest, while GAN-based generators are ranked lowest.

Comparing Evaluators

Kendall τ -b correlation between human and automated evaluators.



Using dataset labels as ground truth



Using human judgements as ground truth

Word-overlap evaluators are positively correlated with the human evaluators in ranking the generators. Surprisingly, **generators that fool discriminative evaluators easily are less likely to confuse human evaluators, and vv.**

Summary of Findings

- Judging the generated text as human-written or machine-generated is a difficult task for humans
- Discriminative evaluators are not as correlated with human judges as word-overlap evaluators: **is adversarial accuracy the optimal objective for realistic text?**
- Training examples must be carefully selected to avoid false-positives when adversarial evaluation is used
- Generators that produce the least diverse samples are easily distinguished by the discriminative evaluators, while they confuse human evaluators the most
- Humans mainly focus on usage of words, expressions, emotions; this may affect the design of objectives for next generation NLG models

Acknowledgements
This work is in part supported by the National Science Foundation under grant numbers 1633370 and 1620319 and by the National Library of Medicine under grant number 2R01LM010681-05.

Paper: https://arxiv.org/pdf/1901.00398.pdf
Code: https://github.com/Crista23/JudgeTheJudges

EMNLP-IJCNLP November 2019