

[Skip to main content](#)> [cs](#) > arXiv:1603.08023

quick links

- [Login](#)
- [Help Pages](#)
- [About](#)

Computer Science > Computation and Language

arXiv:1603.08023 (cs)

[Submitted on 25 Mar 2016 ([v1](#)), last revised 3 Jan 2017 (this version, v2)]

How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation

[Chia-Wei Liu](#), [Ryan Lowe](#), [Iulian V. Serban](#), [Michael Noseworthy](#), [Laurent Charlin](#), [Joelle Pineau](#)[Download PDF](#)

We investigate evaluation metrics for dialogue response generation systems where supervised labels, such as task completion, are not available. Recent works in response generation have adopted metrics from machine translation to compare a model's generated response to a single target response. We show that these metrics correlate very weakly with human judgements in the non-technical Twitter domain, and not at all in the technical Ubuntu domain. We provide quantitative and qualitative results highlighting specific weaknesses in existing metrics, and provide recommendations for future development of better automatic evaluation metrics for dialogue systems.

Comments: First 4 authors had equal contribution. 13 pages, 5 tables, 6 figures. EMNLP 2016

Subjects: **Computation and Language (cs.CL)**; Artificial Intelligence (cs.AI); Machine Learning (cs.LG); Neural and Evolutionary Computing (cs.NE)Cite as: [arXiv:1603.08023](#) [cs.CL]
(or [arXiv:1603.08023v2](#) [cs.CL] for this version)
<https://doi.org/10.48550/arXiv.1603.08023>

[Focus to learn more](#)

arXiv-issued DOI via DataCite

Submission history

From: Ryan Lowe T. [[view email](#)][\[v1\]](#) Fri, 25 Mar 2016 20:32:21 UTC (787 KB)[\[v2\]](#) Tue, 3 Jan 2017 18:28:32 UTC (723 KB)☐ Bibliographic Tools

Bibliographic and Citation Tools

☐ Bibliographic Explorer ToggleBibliographic Explorer ([What is the Explorer?](#))☐ Litmaps ToggleLitmaps ([What is Litmaps?](#))☐ scite.ai Togglescite Smart Citations ([What are Smart Citations?](#))☒ Code, Data, Media

Code, Data and Media Associated with this Article

☐ Links to Code ToggleCatalyzeX Code Finder for Papers ([What is CatalyzeX?](#))☐ DagsHub ToggleDagsHub ([What is DagsHub?](#))☐ Links to Code TogglePapers with Code ([What is Papers with Code?](#))☐ ScienceCast ToggleScienceCast ([What is ScienceCast?](#))☐ Demos

Demos

☐ Replicate ToggleReplicate ([What is Replicate?](#))☐ Spaces ToggleHugging Face Spaces ([What is Spaces?](#))☐ Related Papers

Recommenders and Search Tools

☐ Link to Influence FlowerInfluence Flower ([What are Influence Flowers?](#))☐ Connected Papers ToggleConnected Papers ([What is Connected Papers?](#))☐ Core recommender toggle

CORE Recommender ([What is CORE?](#))

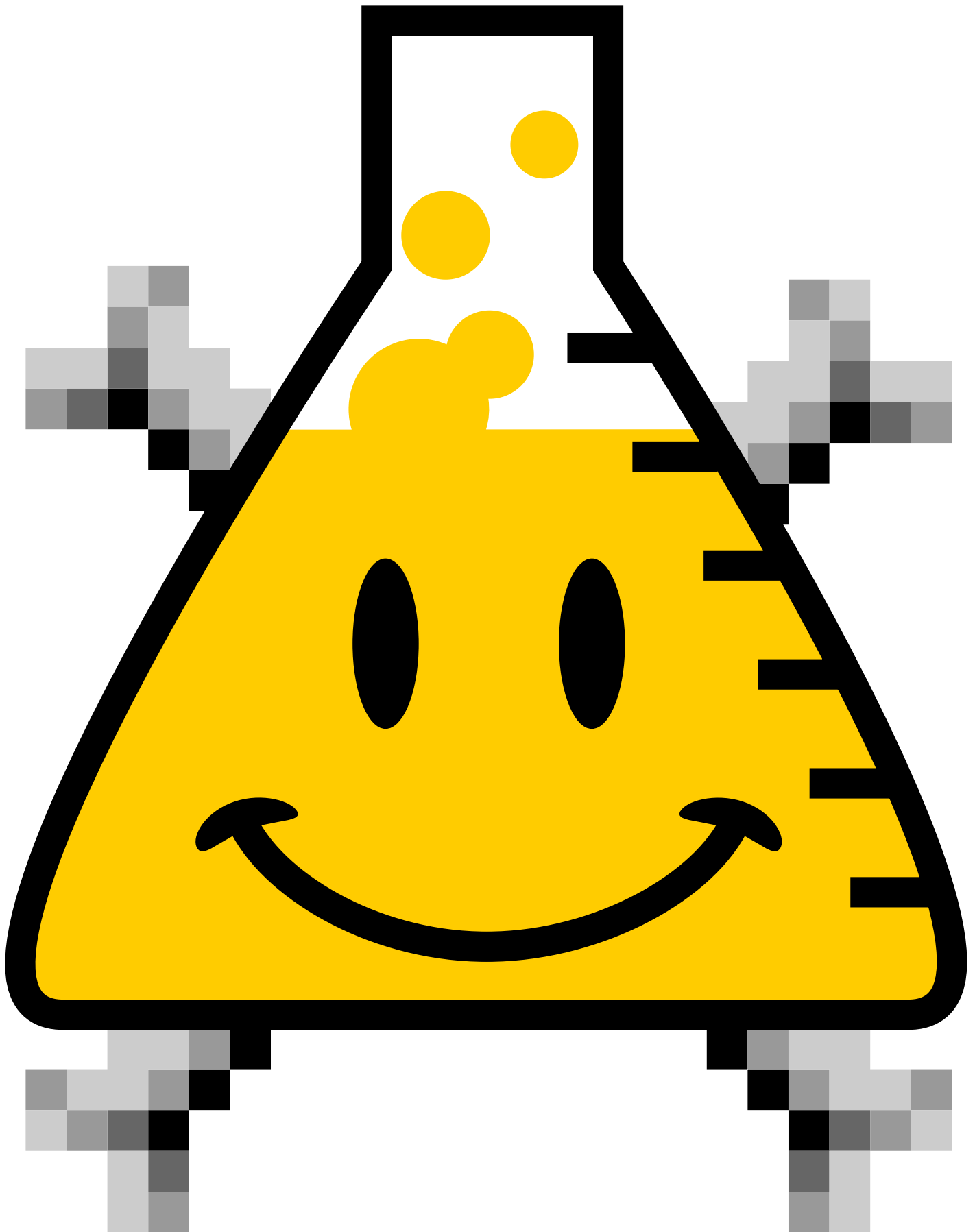
☐ About arXivLabs

arXivLabs: experimental projects with community collaborators

arXivLabs is a framework that allows collaborators to develop and share new arXiv features directly on our website.

Both individuals and organizations that work with arXivLabs have embraced and accepted our values of openness, community, excellence, and user data privacy. arXiv is committed to these values and only works with partners that adhere to them.

Have an idea for a project that will add value for arXiv's community? [Learn more about arXivLabs.](#)



[Which authors of this paper are endorsers?](#) | [Disable MathJax](#) ([What is MathJax?](#))