Evaluation in Natural Language Processing

(Kiela et al., 2021)

Dynabench: Rethinking Benchmarking in NLP

https://dynabench.org/

Content of this work
"We introduce Dynabench, an open-source platform for dynamic dataset creation and model benchmarking. Dynabench runs in a web browser and supports human-and-model-in-the-loop dataset creation: annotators seek to create examples that a target model will misclassify, but that another person will not."

Dynabench

Evaluation of Identifying Language and Script

(Ahmadi et al., 2023)

PALI: A Language Identification Benchmark for Perso-Arabic Scripts

https://github.com/sinaahmadi/PersoArabicLID

Content of this work
"Identifying various languages using such scripts is crucial to language technologies and challenging in low-resource setups. As such, this paper sheds light on the challenges of detecting languages using Perso-Arabic scripts, especially in bilingual communities where "unconventional" writing is practiced. To address this, we use a set of supervised techniques to classify sentences into their languages. Building on these, we also propose a hierarchical model that targets clusters of languages that are more often confused by the classifiers."

Included languages:
Azeri Turkish, Gilaki, Mazanderani, Pashto, Gorani, Northern Kurdish, Central Kurdish, Southern Kurdish, Balochi, Brahui, Kashmiri, Sindhi, Saraiki, Torwali, Punjabi, Persian, Arabic, Urdu, Uyghur

PALI

Evaluation of Multilingual Systems and Models

(Ruder et al., 2021)

XTREME-R: Towards More Chellenging and Nuanced Multilingual Evaluation

Could not find - but maybe not that fitting anyway

Content of this work
"we extend XTREME to XTREME-R, which consists of an improved set of ten natural language understanding tasks, including challenging language-agnostic retrieval tasks, and covers 50 typologically diverse languages. In addition, we provide a massively multilingual diagnostic suite (MULTICHECKLIST) and fine-grained multi-dataset evaluation capabilities through an interactive public leaderboard to gain a better understanding of such models."

XTREME-R

BLEU
↓
BLEURT
METEOR
↓
METEOR UNIVERSAL

Evaluation of Machine Translation

(Papineni et al., 2002)

Bleu: a Method for Automatic Evaluation of Machine Translation

(Sellam et al., 2020)

BLEURT: Learning Robust Metrics for Text Generation

https://github.com/google-research/bleurt

(Lavie and Agarwal, 2007)

METEOR: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments

https://www.cs.cmu.edu/~alavie/METEOR/

(Denkowsi and Lavie, 2014)

Meteor Universal: Language Specific Translation Evaluation for Any Target Language

https://www.cs.cmu.edu/~alavie/METEOR/

(Wang et al., 2018)

GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding

https://gluebenchmark.com/

(Phang et al., 2020)

Jiant 2.0: A software toolkit for research on general-purpose text understanding models

https://github.com/nyu-mll/jiant

(Popović, 2015)

chrF: character n-gram F-score for automatic MT evaluation

https://github.com/m-popovic/chrF

(Rei et al., 2020)

COMET: A Neural Framework for MT Evaluation

https://github.com/Unbabel/COMET

Evaluation in Low-Resource Scenarios

(Riley et al., 2023)

FRMT: A Benchmark for Few-Shot Region-Aware Machine Translation

https://github.com/google-research/google-research/tree/master/frmt

Content of this work
"We explore a setting for MT where unlabeled training data is plentiful for the desired language pair, but only a few parallel examples (0–100, called "exemplars") are annotated for the target varieties."

"We introduce the FRMT dataset for evaluating the quality of few-shot region-aware machine translation. The dataset covers two regions each for Portuguese (Brazil and Portugal) and Mandarin (Mainland and Taiwan)."

Included languages:
Brazilian Portuguese, European Portuguese, Mainland Mandarin, Taiwan Mandarin

FRMT

Evaluation with a Focus on Dialectal Data

(Ziems et al., 2023)

Multi-VALUE: A Framework for Cross-Dialectal English NLP

http://value-nlp.org/

Content of this work
"Still, there does not exist a systematic study on cross-dialectal model performance. We aim to fill this gap, expanding the VernAcular Language Understanding Evaluation (VALUE) framework of Ziems et al. (2022). Where VALUE established a uni-dialectal evaluation harness with 11 perturbation rules, Multi-VALUE now supports multi-dialectal evaluation with 189 different perturbations across 50 English dialects."

Included languages:
50 English dialects

Multi-VALUE

Evaluation concerning Synthetic Data

(Alam et al., 2023)

CoDET: A Benchmark for Contrastive Dialectal Evaluation of Machine Translation

https://github.com/mahfuzibnalam/dialect_mt

Content of this work
"Neural machine translation (NMT) systems exhibit limited robustness in handling source-side linguistic variations. It is intuitive to extend this observation to encompass dialectal variations as well, but the work allowing the community to evaluate MT systems on this dimension is limited. To alleviate this issue, we compile and release CODET, a contrastive dialectal benchmark encompassing 882 different variations from nine different languages."

Included languages:
Arabic, Basque, Bengali, Central Kurdish, Italian, Swiss German, Tigrinya, French, Greek, Yoruba

CODET

chrF

COMET

Use this code to reproduce our baselines. If you want code to use as a starting point for new development, though, we strongly recommend using jiant instead—it's a much more extensive and much better-documented toolkit built around the same goals.

🚨**Update**🚨: As of 2021/10/17, the `jiant` project is no longer being actively maintained.

GLUE → **JIANT**