LiRo: Benchmark and leaderboard for Romanian language tasks — Supplementary

Stefan Dumitrescu
Independent researcher

Petru Rebeja
Alexandru Ioan Cuza University of Iași

Beata Lorincz
Technical University of Cluj-Napoca

Mihaela Gaman
University of Bucharest

Andrei-Marius Avram
Politehnica University of Bucharest

Mihai Ilie
Independent researcher

Andrei Pruteanu
Independent researcher

Adriana Stan
Technical University of Cluj-Napoca

Lorena Rosia
Deloitte

Cristina Iacobescu
Deloitte

Luciana Morogan
Military Technical Academy

George-Andrei Dima
Politehnica University of Bucharest

Gabriel Marchidan
Feel IT Services

Traian Rebedea
Politehnica University of Bucharest

Madalina Chitez
West University of Timisoara

Dani Yogatama
DeepMind

Sebastian Ruder
DeepMind

Radu Tudor Ionescu
University of Bucharest

Razvan Pascanu
DeepMind

Viorica Patraucean
DeepMind
viorica@google.com

1 *LiRo*

LiRo is a benchmark for evaluating models on Romanian language tasks, inspired by similar benchmarks for other languages, e.g. GLUE for English [5], KLEJ for Polish [4], KLUE for Korean [3]. At the moment of writing this paper, the benchmark consists of nine standard tasks (text classification, named entity recognition, machine translation, sentiment analysis, POS tagging, dependency parsing, language modelling, question-answering, and semantic textual similarity), and an extra task of Romanian embeddings gender debiasing, to address the growing concerns related to gender bias in language models. The platform exposes per-task leaderboards populated with baseline results for each task. These tasks rely on existing datasets or newly created ones. In particular, we created three new datasets: one from Romanian Wikipedia and two by translating the Semantic Textual Similarity (STS) benchmark and the Cross-lingual Question Answering Dataset (XQuAD) into Romanian. We include below summarised documentation for each dataset. Full details about the dataset collection and content are included in the main paper.

1.1 Intended Uses

LiRo is intended for researchers in natural language processing, machine learning, and related fields to develop novel methods for Romanian language understanding and cross-lingual studies.

35th Conference on Neural Information Processing Systems (NeurIPS 2021) Track on Datasets and Benchmarks.

1.2 Hosting and Maintenance Plan

LiRo is hosted and version-tracked via GitHub. It will be permanently available under the link `https://lirobenchmark.github.io/`. The download link of all the datasets can be found under each task listed on the website.

LiRo is a an open-source community-driven initiative. We are committed to maintain and actively develop *LiRo* for a minimum of 5 years since the publication date. We plan to expand *LiRo* with new learning tasks and datasets. We welcome external contributors. Contact us at `liro.benchmark@gmail.com`.

1.3 Licensing

LiRo is a collection of open-source datasets, each having its own license agreement; see details below and on *LiRo* webpage[1].

1.4 Author Statement

We, the authors, will bear all responsibility in case of violation of rights.

2 Data Statements for RO-STS Dataset

The Romanian Semantic Textual Similarity Dataset RO-STS is the Romanian translation of the English STS dataset [2]. It contains 8628 Romanian sentence pairs with their similarity scores. The sentence pairs belong to categories such as news headlines, image captions, and user forums.

2.1 Curation Rationale

RO-STS dataset is the Romanian version of the Semantic Textual Similarity Dataset [2]. It is a high-quality dataset created with the scope of expanding the available resources for Romanian language.

The dataset is suitable as: (i) a textual similarity dataset, and (ii) a parallel Romanian-English corpus, having the dev/train/test splits identical to the original STS corpus splits.

2.2 Language Variety

The dataset was created by translating text from English language, for which, at the moment of writing, there is no information available on variety but at least United States English (en-US) is included.

The translation was performed by native Romanian speakers originating from various regions of Romania; as such the Romanian translations are under the ro-RO variety.

2.3 Speaker Demographic

Since the RO-STS dataset is a translation of the original dataset, the speaker information is the same. However, at the moment of writing, there is no speaker information available for the original STS corpus.

2.4 Annotator Demographic

The translation of the STS corpus was performed by 13 volunteers, all of which are native Romanian speakers, are fluent in English, and are residents of Romania and United Kingdom. The volunteers range in age from 20 to 40 years, and include 6 men and 7 women.

[1] `https://lirobenchmark.github.io/terms-and-conditions`

2.5 Speech Situation

Both the original and the translated texts represent spontaneous writing language, with no limit on the number of characters. The Romanian translations were created in the first half of 2021.

2.6 Text Characteristics

The samples of the dataset pertain to various genres such as news headlines, captions of images, user forums etc. There is no predominant topic in the dataset, and all of the samples use a single modality — text.

2.7 Hosting and Maintenance Plan

RO-STS dataset is hosted on Github at the following URL: `https://github.com/dumitrescustefan/RO-STS`. The README page provides download links for: (i) the whole dataset as a single zip file or, (ii) three separate files for development, training, and testing.

2.8 Licensing

The dataset is licensed under CC BY-SA 4.0 license. This license allows the users to share and adapt the dataset as long as proper attribution is provided, and the adapted dataset is distributed under the same license. The full text of the license can be consulted at `https://creativecommons.org/licenses/by-sa/4.0/`.

2.9 Author Statement

The authors will bear all responsibility in case of violation of rights.

2.10 Provenance Appendix

RO-STS dataset is built from STS bechmark dataset, which is available at `https://ixa2.si.ehu.eus/stswiki/index.php/STSbenchmark`.

3 Data Statements for XQuAD-ro Dataset

XQuAD-ro is the Romanian component of the XQuAD dataset [1]. XQuAD (Cross-lingual Question Answering Dataset) is a benchmark dataset for evaluating cross-lingual question answering performance. The dataset consists of a subset of 240 paragraphs and 1190 question-answer pairs from the development set of SQuAD v1.1 together with their professional translations into Spanish, German, Greek, Russian, Turkish, Arabic, Vietnamese, Thai, Chinese, Hindi and (our newly-added) Romanian. The dataset is entirely parallel across all languages.

3.1 Curation Rationale

XQuAD-ro dataset is created with the intentions of extending the original XQuAD dataset with data for Romanian language, and of adding resources for the Romanian NLP community.

3.2 Language Variety

XQuAD-ro is a translation of the original XQuAD dataset for which, at the moment of writing, there is no information available on language variety, but at least the United States English (en-US) is included.

The translation was performed by professional translators; as such the language variety of XQuAD-ro is ro-RO.

3.3 Speaker Demographic

Since XQuAD-ro is a translation of the original dataset, the speaker information is the same, but at the moment of writing, there is no speaker demographic information available for XQuAD dataset.

3.4 Annotator Demographic

There is no demographic data available for the translators that performed English to Romanian translation of texts included in XQuAD-ro.

3.5 Speech Situation

Both the original XQuAD dataset and XQuAD-ro consist of texts that represent spontaneous writing language, with no limit on the number of characters. The questions tend to be short while the context for the questions is longer.

3.6 Text Characteristics

XQuAD-ro dataset contains data in a single modality — text. The questions from the dataset are general knowledge questions and no single predominant topic emerges from the questions.

3.7 Hosting and Maintenance Plan

The dataset is available for download via GitHub at `https://github.com/deepmind/xquad`.

3.8 Licensing

The dataset is licensed under CC BY-SA 4.0 license. This license allows the users to share and adapt the dataset as long as proper attribution is provided, and the adapted dataset is distributed under the same license. The full text of the license can be consulted at `https://creativecommons.org/licenses/by-sa/4.0/`.

3.9 Author Statement

The authors will bear all responsibility in case of violation of rights.

3.10 Provenance Appendix

XQuAD-ro dataset is a translation of XQuAD dataset. The Romanian version, alongside the versions in other languages are hosted on Github at `https://github.com/deepmind/xquad`.

4 Data Statements for Wiki-ro Dataset

The Wiki-ro corpus consists of cleaned text extracted from the Romanian Wikipedia. This corpus is meant to test the capacity of a language model (LM) by measuring its perplexity on the test set after being trained solely on the training data. LMs pretrained on external data can be tested as well, but will have the flag marking their out-of-domain pretraining. The dataset is divided into train, validation, and test splits, always making sure that a document is entirely included in a single split. The train, validation, and test sets have 2.1M lines and 44M words, 14K lines and 276K words, and 16K lines and 327K words, respectively.

4.1 Curation Rationale

As with the other datasets from this paper, Wiki-ro dataset was created to enrich the landscape of Romanian NLP resources. It is intended for evaluation of language models for Romanian text.

4.2 Language Variety

There is no exact data on language variety from Wikipedia for texts in Romanian language but at least mainstream Romanian (ro-RO) is present.

4.3 Speaker Demographic

No demographic data is available on creators of the Romanian Wikipedia articles.

4.4 Annotator Demographic

Not applicable — the data in Wiki-ro dataset is not annotated.

4.5 Speech Situation

The Wiki-ro dataset contains articles that were published on Romanian Wikipedia since July 2003 and June 2020. The articles represent topic-based written language with no limit on number of characters and a single topic per article. The intended audience of articles is general audience.

4.6 Text Characteristics

Although the original Romanian Wikipedia articles may contain both text and media resources, the data in Wiki-ro is of single modality, namely text. The articles pertain to various topics, with one topic per article.

4.7 Hosting and Maintenance Plan

The dataset is available for download via GitHub at `https://github.com/dumitrescustefan/wiki-ro`.

4.8 Licensing

Wiki-ro is released under the MIT license. This license allows usage of the dataset for commercial purposes, distribution and modification of the data as long as a copy of the license and copyright notice are included with the shared material. The full text of the MIT license can be consulted at `https://choosealicense.com/licenses/mit/`.

4.9 Author Statement

The authors will bear all responsibility in case of violation of rights.

4.10 Provenance Appendix

This is a new dataset built from a database dump of Romanian Wikipedia.

References

[1] Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the Cross-lingual Transferability of Monolingual Representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637. Association for Computational Linguistics.

[2] Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada. Association for Computational Linguistics.

[3] Sungjoon Park, Jihyung Moon, Sungdong Kim, Won-Ik Cho, Jiyoon Han, Jangwon Park, Chisung Song, Junseong Kim, Yongsook Song, Tae Hwan Oh, Joohong Lee, Juhyun Oh, Sungwon Lyu, Younghoon Jeong, Inkwon Lee, Sangwoo Seo, Dongjun Lee, Hyunwoo J. Kim, Myeonghwa Lee, Seongbo Jang, Seungwon Do, Sunkyoung Kim, Kyungtae Lim, Jongwon Lee, Kyumin Park, Jamin Shin, Seonghyun Kim, Eunjeong Lucy Park, Alice Oh, Jung-Woo Ha, and Kyunghyun Cho. 2021. KLUE: korean language understanding evaluation. *CoRR*, abs/2105.09680.

[4] Piotr Rybak, Robert Mroczkowski, Janusz Tracz, and Ireneusz Gawlik. 2020. KLEJ: Comprehensive Benchmark for Polish Language Understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 1191–1201. Association for Computational Linguistics.

[5] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.