

Evaluation of Large Language Models in Multilingual Settings

Maite Heredia Arribas

Universidad del País Vasco / Euskal Herriko Unibertsitatea, Barrio Sarriena, 48940 Leioa, Bizkaia

Abstract

Evaluation is an essential step in developing NLP models, that is gaining a lot of relevance alongside the need for more sophisticated metrics to account for the growth of LLMs (Large Language Models). These models have demonstrated very competitive results in a broad range of tasks and are showing new, broader skills that were previously unthinkable. Our research focuses on evaluating the multilingual capabilities of LLMs, specially for the creation of new benchmarks. They have been reported to be able to transfer tasks among languages and have great potential for adapting the available resources for less-resourced languages. We aim to design new benchmarks for these languages to help reduce the gap between the leading languages in NLP (and specifically English) and languages that are still lacking behind in resources and research. More concrete objectives include defining essential evaluation metrics across languages lacking benchmarks, comparing benchmark creation methods (automatic translation, human translation, and creation from scratch), and evaluating multilingual LLM performance with fine-tuning and zero-shot techniques. Initial efforts have been focused on creating benchmarks for common sense reasoning and code-switched text. Future research will expand dataset creation efforts and explore fine-tuning strategies to enhance multilingual LLM performance.

Keywords

evaluation, Large Language Models, multilingualism, low-resource languages, dataset creation, cross-linguistic evaluation


1. Reason for the proposed research

Most modern benchmarks allow us to broadly categorize the performance of LLMs, but they generally fail in three areas: 1) they do not widely evaluate languages other than English, and often overlook low-resource languages, which is the case of Iberian languages like Basque, Catalan and Galician [1]; 2) they are opportunistic, in that they are made from a collection of tasks which were already available, which are oftentimes not what we would naturally use LLMs for, and models that obtain remarkably good results in popular benchmarks have been proven to fail on simple test cases [2]; and 3) they generally only measure a single performance metric, i.e., accuracy, unsuited to accurately measure generative models' capabilities, since they are not able to capture semantic meaning [3]. To properly evaluate LLMs on languages other than English, we therefore need to devise new benchmarks and metrics which will take into account more than just a single performance metric and will be culturally and linguistically diverse.

Doctoral Symposium on Natural Language Processing, 26 September 2024, Valladolid, Spain.

✉ maite.heredia@ehu.eus (M. Heredia Arribas)

ORCID 0009-0005-6719-5433 (M. Heredia Arribas)

 © 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

Therefore, the main objective of this PhD project is to explore novel methods for the creation of relevant benchmarks for LLMs, specifically with and for multilingual settings. Using these benchmarks, we will be able to explore multilingual capabilities of LLMs among high- and low-resource languages, their ability to transfer tasks to different languages and their potential to devise new datasets.

2. Background and related work

Large Language Models (LLMs) have recently led to a seemingly unprecedented level of performance in Natural Language Processing (NLP)[4, 5]. These models are trained on vast amounts of text, with the goal of learning representations of language which can then be transferred to new scenarios with a minimum amount of fine-tuning. While certain models trained by private companies, such as ChatGPT, have perhaps had the most media coverage, academic models (PaLM, LLaMA) have also made progress.

Due to their black box nature, LLMs are difficult to interpret and understand. The knowledge acquired by the model during training is distributed among hundreds of millions of parameters, and therefore it is very difficult to interpret the output produced by the models, or the underlying reasons that steer the model to generate certain text. Moreover, large scale models have shown unexpected capabilities that have surprised the research community, as they were not specifically designed to acquire them. These so-called "emergent abilities", the capacity of the models to resolve tasks for which they have not been previously trained, just by providing them with very few training examples [6], have turned out to be one of the most important characteristics of large neural models that allow their deployment in many NLP applications and domains.

While the progress made recently is undoubtable, our understanding of what this progress really means is limited by our evaluation methods, metrics, and benchmarks that are currently available. Neural models have improved to the point where they can often no longer be distinguished based on the surface-level features that older metrics rely on [7], and there is a growing need to devise evaluation strategies that measure not only the progress of the models, but that also help us to understand the properties of large language models, their capabilities, limitations, and risks. This will help in turn overcoming various shortcomings of current LLM approaches that are critical for widespread adoption and that have, so far, not been successfully addressed or solved. We focus on a specific capability of LLMs, multilingualism.

Several new benchmarks have been proposed for evaluating LLMs. To create MMLU, [8], for example, collect over 15,000 multiple choice questions in English – taken from American GRE and medical licensing exam preparation courses – and divided into 57 different topics, broadly grouped into STEM, Humanities, Social Sciences, and Other. [9] propose BIG-Bench with the purpose of creating a benchmark that would be more difficult and last longer than previous benchmarks. They take a crowd-sourcing strategy and collaboratively create 204 tasks, some of which are also available in languages other than English. [10] propose HELM (Holistic Evaluation of Language Models), which instead takes a top-down approach, spelling out the scenarios they evaluate on and which are currently missing. Like other proposed benchmarks, they limit themselves to collecting already available resources for several tasks – question answering, information retrieval, summarization, sentiment analysis, toxicity, and text

classification – but furthermore incorporate a number of evaluation metrics besides accuracy (bias, fairness, efficiency, robustness, toxicity, and calibration).

All the datasets described so far are only available for English. While there exist several LLMs for many languages other than English, the evaluation benchmarks on those languages clearly lag behind. In the Spanish State, the NEL project founded by the national government (https://www.boe.es/diario_boe/txt.php?id=BOE-A-2022-18816) seeks to build a new generation of LLMs for Iberic languages that compete in performance with English LLMs [11, 12]. The project emphasizes the need to develop and research on methods to build benchmarks to evaluate the capacities of LLMs for said languages, an objective that is fully aligned with this thesis work.

Multilingual models are designed to handle multiple languages simultaneously, leveraging the shared information across languages to improve performance on individual tasks. These models, such as mBERT, XLM-R, and mT5, are trained on multilingual corpora and have demonstrated significant improvements in understanding and generating text in various languages, including those with limited training data [13]. The capability of multilingual LLMs to transfer knowledge across languages makes them particularly valuable for less-resourced languages, where annotated data is scarce. Our research aims to address these challenges by creating new benchmarks specifically tailored for evaluating the multilingual capabilities of LLMs, thereby contributing to the advancement of NLP for a broader range of languages.

3. Description of the proposed research

3.1. Initial proposal

To enable the evaluation of multilingual LLMs, we first defined the following specific objectives:

- Define a set of variables / metrics that are relevant for evaluating LLMs in languages that do not currently have an evaluation benchmark. These could include (but are not limited to):
 - Grammatical abilities: How well does the model handle complex morpho-syntax, coreference, etc.?
 - Logic abilities: Can the model reason over complex logical structures?
 - Common sense: What common sense knowledge is encoded in the LLM?
 - Code generation abilities: Can the model be used to generate runnable code?
 - Truthfulness: How likely is it that the LLM presents fabricated information as truth?
 - Bias: What kinds of biases does the model have?
 - Toxicity: Under what circumstances does the model produce toxic output?
 - Non-standard language: How well can the model operate with text from social media, non-standard language varieties, or code-switching?

An extensive review of the literature will allow us to more clearly define what are the most relevant variables and metrics and find the gaps in current resources and research to try to help in filling them.

- In order to create a benchmark, compare automatic translation of benchmarks, human translation, and creation from scratch.

- Compare the performance of multilingual / monolingual models using different fine-tuning and prompting techniques, and use this information to improve model weaknesses.

3.2. Current progress

So far we have tackled the creation of benchmarks for common sense, specifically for the NLI task¹ [14]; and non-standard language, more specifically social media text that includes code-switching [15]. We have centered our work around the Basque language, but we are aiming to expand our research to more languages.

Our work so far has allowed us to explore the impact of different methods of creating evaluation sets [16] - machine translation, human translation, from scratch - and we have been able to gather some initial conclusions. More specifically, we have corroborated that machine translation is a useful resource to create datasets, more so if the goal is model comparison, but, when available, human post-edition and even creation from scratch are preferable and can more accurately assess the capabilities of models, which underscores the effectiveness of certain approaches in preserving linguistic nuances and ensuring benchmark quality. Apart from creating our datasets and making our datasets publicly available, we have also performed a first batch of experiments to test different multi- and cross-lingual strategies to leverage the existing resources in English and other high-resource languages for languages with fewer resources. In agreement with other researches' results [13], our experiments show that using multilingual LLMs with strategies like zero-shot cross-lingual transfer, translate-train or zero-shot prompting can be feasible alternatives in scenarios where there are not enough resources for more standard approaches.

3.3. Next steps

We would like to continue with our work so far by developing more datasets, using the knowledge that we have gathered about the importance of human supervision, and testing whether our findings hold true in different tasks and settings. Currently, we are also exploring the possibility of creating datasets by making use of generative autoregressive LLMs, and would like to continue researching cross-lingual transfer.

4. Methodology and proposed experiments

As an overview, our proposed experiments will mainly consist on researching the most efficient methods for creating new benchmarks, that involve the acquisition of resources and their processing, translation and/or annotation. To assure the effectiveness of these resources, it will be essential to perform experiments with them, that may involve fine-tuning and testing LLMs. We will publicly release the created benchmarks, as well as an environment to evaluate LLMs and share these with the scientific community. In this way we will provide comparable results and we will gain insight about improvements. Although some of the ideas may not provide the desired results, we will fulfill all the objectives and get feedback from the results. All the true hypotheses will be shared in major peer-reviewed conferences.

¹<https://github.com/hitz-zentroa/xnli-eu>

4.1. Main Hypotheses

Here, we present the four main hypotheses that guide our work. These hypotheses are further elaborated and expanded as we progress in our research, incorporating new insights and developments.

H0: Transferring English benchmarks to other languages. The benchmarks and metrics that have been developed for English are not suitable to measure the performance in a lot of other languages, especially low-resourced languages, and have to be adapted and improved.

H1: Human intervention. Human intervention is essential to create reliable benchmarks for languages other than English, especially low-resourced ones.

H2: Cultural impact. Benchmarks devised for English are culturally bounded to the Anglo-sphere. Therefore, the development of resources for different languages should not reproduce these cultural biases, but rather try to adapt to the culture of each language.

H3: Non-standard language. A great deal of well-known and widely used benchmarks and metrics are not representative of real language production, but rather a standardized form of language.

4.2. Research Tasks (RT), Questions (RQ)

RT0. Prepare the research scenario. The initial task involves defining the variables of interest, finding metrics to correctly measure these variables, and collecting relevant datasets which are available in English and other high-resource languages. We will perform initial experiments in English to determine the feasibility of our metrics and tests RQ0.A) What current datasets are available in English? RQ0.B) Which variables (performance, bias, toxicity, etc) do these datasets measure? RQ0.C) Do the current metrics correctly capture the most important variables?

RT1. Explore the most appropriate method for creating successful LLM benchmarks for low-resourced languages. In this task, we will compare the strengths and weaknesses of creating new benchmarks through translation-based transfer or by creating the resources from scratch. RQ1.A) Can automatic translation create robust benchmarks in new languages? RQ1.B) Can human translation create robust benchmarks in new languages? RQ1.C) What culturally specific artifacts are lost in translation? RQ1.D) How does translation-based benchmarking differ from creating new benchmarks from scratch?

RT2. Determine the correlations between the performance of LLMs on standardized tests as benchmarks and their performance on other tasks. RQ2.A) How does model performance on available standardized tests correlate with performance on NLP-style tasks? RQ2.B) Does the performance on the standardized tests or NLP tasks correlate with the ability of models to perform useful functions, e.g. write a formal email conditioned on some information?

RT3. Compare monolingual and multilingual models, including English models. RQ3.A) What relative strengths/weaknesses do multilingual and monolingual models have? RQ3.B) How does the monolingual performance of these models compare to English models, such as GPT4? RQ3.C) How does finetuning, instruction tuning, and prompt engineering change the performance of these models?

4.3. Research schedule yearly

First year: In the first year we will mainly focus on tasks related to RT0. The goal is to prepare the research scenario so we will need to define the main variables of interest and collect available English datasets. We foresee the following tasks:

- We will gather the basic resources, such as a collection of the main datasets used to evaluate LLMs in English, as well as any standardized tests available in Iberian languages, which help answer RQ0.A and RQ0.B.
- We will create a taxonomy of the variables that are measured in these datasets and their corresponding metrics, in order to answer RQ0B.
- We will design and run experiments on state-of-the-art English LLMs on the datasets from the previous tasks to answer RQ0.C.
- We aim to submit the answers of RQs to a top journal or conference.

Second year: During this year we will mainly focus on tasks RT1 and RT2. For that, we first will experiment with methods to create a comprehensive benchmark for Iberian LLMs, comparing translation methods and in-language annotation. In the second part of the year, we will start working in parallel with the comparison of standardized testing with NLP tasks. For this year, we foresee the following tasks:

- We will use automatic translation to transfer the available English resources to Iberian languages and perform an analysis of the resulting translated datasets, paying attention to what errors are introduced, what topics are included, and LLM performance. This task will help answer RQ1.A.
- From the analysis of the previous step, we will choose a subset of data to translate via human translators. We will perform a similar analysis of errors, topics, and LLM performance. This task will help answer RQ1.B.
- Finally, we will perform an annotation project on a subset of the datasets used in human translation. The goal will be to create similar datasets, but localized and annotated by native speakers. We will compare the distribution of topics, as well as LLM performance on this data with the translated versions. This task will allow us to answer both RQ1.C and RQ1.D
- To answer RQ2A, we will compare LLM performance on our benchmark and available standardized exams in Iberian languages. We will perform a regression analysis of performance improvements on the standardized exams and other tasks to determine what relation exists.

- For RQ2.B, we will need to determine a small set of tasks that users of LLMs would be interested in, which we would gather via a small community survey. Once we have the results of the survey, we can test the models on these tasks and perform a similar analysis as we did to answer RQ2A.
- Note that due to the great number of tasks planned for the second year, we probably need to postpone some tasks to the third year.

Third year: During the third year, we will focus on tasks related to RT3. Our objective is to compare the strengths of monolingual Iberian LLMs, multilingual Iberian LLMs, massively multilingual LLMs, and monolingual English LLMs.

- In order to answer RQ3.A, we first compare available monolingual and multilingual LLMs on our benchmark.
- Compare these results to the results of English LLMs to answer RQ3.B.
- Given the insights from the previous experiments, propose finetuning, instruction tuning, and prompting methods to improve the weaknesses of monolingual models (RQ3.C).
- Depending on the answers of RQ3s, we will submit our work to a top conference.

Fourth year: The most interesting conclusions obtained from previous years will be taken and rounded off through new experiments in the first months of the year. Then the thesis will be written and the thesis defense preparation will be done. The following tasks are planned for this purpose:

- Finish tasks and experiments of previous years.
- Submitting related research to a top journal or conference.
- Write-up of the PhD thesis.

5. Specific issues of research to be discussed

Our research is mainly focused on the creation and evaluation of reliable benchmarks for low-resource languages, which is undoubtedly one critical bottleneck in the development of NLP applications. Creating new datasets is always a time-consuming task, specially when working with models that need large amounts of training data. Most new datasets are either opportunistic, in the sense that they are stemmed from already existing data that has been pre-annotated for different purposes, or are created through crowd-sourcing, that can result in lower quality annotations (and, sometimes, unethical work practices [17]). These can be valid approaches to the creation of datasets, but we argue that for benchmarks to be more useful and more accurately measure the capabilities of models, they should ideally be more carefully designed, linguistically motivated, and annotated by professionals of different areas of expertise, depending on the task. These considerations make our work slower and significantly more challenging in different steps of the workflow: the design of the process, data collecting and, most notably, annotating.

On the other hand, it is worth mentioning that one of the possible lines of research of this thesis is to evaluate ethical aspects of large language models, including biases and harmfulness. These are sensitive topics that will be approached with the necessary care and consideration.

References

- [1] P. Joshi, S. Santy, A. Budhiraja, K. Bali, M. Choudhury, The state and fate of linguistic diversity and inclusion in the NLP world, in: D. Jurafsky, J. Chai, N. Schluter, J. Tetreault (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 6282–6293. URL: <https://aclanthology.org/2020.acl-main.560>. doi:10.18653/v1/2020.acl-main.560.
- [2] S. R. Bowman, G. Dahl, What will it take to fix benchmarking in natural language understanding?, in: K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, Y. Zhou (Eds.), Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Online, 2021, pp. 4843–4855. URL: <https://aclanthology.org/2021.naacl-main.385>. doi:10.18653/v1/2021.naacl-main.385.
- [3] T. Linzen, How can we accelerate progress towards human-like linguistic generalization?, in: D. Jurafsky, J. Chai, N. Schluter, J. Tetreault (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 5210–5217. URL: <https://aclanthology.org/2020.acl-main.465>. doi:10.18653/v1/2020.acl-main.465.
- [4] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language models are few-shot learners, in: H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin (Eds.), Advances in Neural Information Processing Systems, volume 33, Curran Associates, Inc., 2020, pp. 1877–1901. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfc4967418bfb8ac142f64a-Paper.pdf.
- [5] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann, P. Schuh, K. Shi, S. Tsvyashchenko, J. Maynez, A. Rao, P. Barnes, Y. Tay, N. Shazeer, V. Prabhakaran, E. Reif, N. Du, B. Hutchinson, R. Pope, J. Bradbury, J. Austin, M. Isard, G. Gur-Ari, P. Yin, T. Duke, A. Levskaya, S. Ghemawat, S. Dev, H. Michalewski, X. Garcia, V. Misra, K. Robinson, L. Fedus, D. Zhou, D. Ippolito, D. Luan, H. Lim, B. Zoph, A. Spiridonov, R. Sepassi, D. Dohan, S. Agrawal, M. Omernick, A. M. Dai, T. S. Pillai, M. Pellat, A. Lewkowycz, E. Moreira, R. Child, O. Polozov, K. Lee, Z. Zhou, X. Wang, B. Saeta, M. Diaz, O. Firat, M. Catasta, J. Wei, K. Meier-Hellstern, D. Eck, J. Dean, S. Petrov, N. Fiedel, Palm: Scaling language modeling with pathways, 2022. [arXiv:2204.02311](https://arxiv.org/abs/2204.02311).
- [6] J. Wei, Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler, E. H. Chi, T. Hashimoto, O. Vinyals, P. Liang, J. Dean, W. Fedus, Emergent abilities of large language models, 2022. [arXiv:2206.07682](https://arxiv.org/abs/2206.07682).
- [7] S. Gehrmann, E. Clark, T. Sellam, Repairing the cracked foundation: A survey of obstacles in evaluation practices for generated text, 2022. [arXiv:2202.06935](https://arxiv.org/abs/2202.06935).
- [8] D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, J. Steinhardt, Measuring massive multitask language understanding, Proceedings of the International Conference

on Learning Representations (ICLR) (2021).

- [9] BIG-Bench authors, Beyond the imitation game: Quantifying and extrapolating the capabilities of language models, 2023. [arXiv:2206.04615](https://arxiv.org/abs/2206.04615).
- [10] P. Liang, R. Bommasani, T. Lee, D. Tsipras, D. Soylu, M. Yasunaga, Y. Zhang, D. Narayanan, Y. Wu, A. Kumar, B. Newman, B. Yuan, B. Yan, C. Zhang, C. Cosgrove, C. D. Manning, C. Ré, D. Acosta-Navas, D. A. Hudson, E. Zelikman, E. Durmus, F. Ladhak, F. Rong, H. Ren, H. Yao, J. Wang, K. Santhanam, L. Orr, L. Zheng, M. Yuksekgonul, M. Suzgun, N. Kim, N. Guha, N. Chatterji, O. Khattab, P. Henderson, Q. Huang, R. Chi, S. M. Xie, S. Santurkar, S. Ganguli, T. Hashimoto, T. Icard, T. Zhang, V. Chaudhary, W. Wang, X. Li, Y. Mai, Y. Zhang, Y. Koreeda, Holistic evaluation of language models, 2023. [arXiv:2211.09110](https://arxiv.org/abs/2211.09110).
- [11] S. Da Dalt, J. Llop, I. Baucells, M. Pamies, Y. Xu, A. Gonzalez-Agirre, M. Villegas, FLOR: On the effectiveness of language adaptation, in: N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, N. Xue (Eds.), Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), ELRA and ICCL, Torino, Italia, 2024, pp. 7377–7388. URL: <https://aclanthology.org/2024.lrec-main.650>.
- [12] J. Etxaniz, O. Sainz, N. Perez, I. Aldabe, G. Rigau, E. Agirre, A. Ormazabal, M. Artetxe, A. Soroa, Latxa: An open language model and evaluation suite for basque, 2024. URL: <https://arxiv.org/abs/2403.20266>. [arXiv:2403.20266](https://arxiv.org/abs/2403.20266).
- [13] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, 2020. [arXiv:1911.02116](https://arxiv.org/abs/1911.02116).
- [14] A. Conneau, R. Rinott, G. Lample, A. Williams, S. Bowman, H. Schwenk, V. Stoyanov, XNLI: Evaluating cross-lingual sentence representations, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 2475–2485. URL: <https://aclanthology.org/D18-1269>. doi:10.18653/v1/D18-1269.
- [15] G. Winata, A. F. Aji, Z. X. Yong, T. Solorio, The decades progress on code-switching research in NLP: A systematic survey on trends and challenges, in: A. Rogers, J. Boyd-Graber, N. Okazaki (Eds.), Findings of the Association for Computational Linguistics: ACL 2023, Association for Computational Linguistics, Toronto, Canada, 2023, pp. 2936–2978. URL: <https://aclanthology.org/2023.findings-acl.185>. doi:10.18653/v1/2023.findings-acl.185.
- [16] M. Artetxe, G. Labaka, E. Agirre, Translation artifacts in cross-lingual transfer learning, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, 2020, pp. 7674–7684. URL: <https://aclanthology.org/2020.emnlp-main.618>. doi:10.18653/v1/2020.emnlp-main.618.
- [17] B. Shmueli, J. Fell, S. Ray, L.-W. Ku, Beyond fair pay: Ethical implications of NLP crowdsourcing, in: K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, Y. Zhou (Eds.), Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Online, 2021, pp. 3758–3769. URL: <https://aclanthology.org/2021.naacl-main.295>. doi:10.18653/v1/2021.naacl-main.295.