On the impact of Language Adaptation for Large Language Models: A case study for the Italian language using only open resources

Pierpaolo Basile[1], Pierluigi Cassotti[1], Marco Polignano[1], Lucia Siciliani[1] and Giovanni Semeraro[1]

[1]University of Bari Aldo Moro, Department of Computer Science, via E. Orabona, 70125, Bari, Italy

Abstract

The BLOOM Large Language Model is a cutting-edge open linguistic model developed to provide computers with natural language understanding skills. Despite its remarkable capabilities in understanding natural language by capturing intricate contextual relationships, the BLOOM model exhibits a notable limitation concerning the number of included languages. In fact, Italian is not included among the languages supported by the model making the usage of the model challenging in this context. Within this study, using an open science philosophy, we explore different *Language Adaptation* strategies for the BLOOM model and assess its zero-shot prompting performance across two different downstream classification tasks over EVALITA datasets. It has been observed that language adaptation followed by instruction-based fine-tuning is shown to be effective in correctly addressing a task never seen by the model in a new language learned on a few examples of data.

Keywords

Natural Language Processing, Language Adaptation, Large Language Model

1. Introduction

As language diversity becomes increasingly important in the digital age, the capability of a Natural Language Understanding model to handle a wide array of languages gains significance. Large Language Models (LLMs) have emerged as excellent approaches for comprehending, generating, and manipulating human language with unprecedented accuracy and fluency [1].

They can grasp nuances, idioms, and even ambiguous phrases, enabling more accurate sentiment analysis, question answering, and information retrieval tasks. This enhanced understanding contributes to more effective communication between humans and machines, fostering seamless interactions across various applications. LLMs possess remarkable generalization capabilities, allowing them to perform well on tasks they were not explicitly trained for, also in a multilingual fashion. Among the largest and most effective Large Language Models can be found BLOOM [2], a 176B-parameter open-access language model designed and built thanks to the collaboration of hundreds of researchers. BLOOM is a decoder-only Transformer language model that was trained on a large corpus comprising hundreds of sources in 46 natural and 13 programming languages, culminating in a comprehensive dataset that spans 59 languages in total. Nevertheless, it excludes some of the world's most widely spoken languages, including Russian, Korean, and Italian, raising the need for a more inclusive linguistic approach. Training an effective LLM focused solely on a particular language is a prohibitive challenge, given the substantial volumes of data and resources required for such a task. At the same time, tackling downstream tasks in a specific language effectively necessitates a model with a comprehensive understanding of that language.

Our hypothesis focuses on the Language Adaptation methodology, which is particularly fascinating for addressing the challenge of transferring knowledge from a pre-trained Language Model (LM) to a specific application language. In this context, we aim to adapt BLOOM models to work with a new language, such as Italian, using only a limited sample size, i.e., 100,000 samples.

Indeed, we evaluated the adapted models after a phase of instruction-based fine-tuning on two different classification tasks using Italian data. Our experiments demonstrate that the Language Adaptation process improves the zero-shot ability of the model if executed for the same language of the evaluating data. One of the most important aims of our work is the development of all the methodologies using an open-science approach without using private data created or elaborated by no open-source tools. In addition to this, all data and models used in this work are under an open-source license, reflecting our

CLiC-it 2023: 9th Italian Conference on Computational Linguistics, Nov 30 — Dec 02, 2023, Venice, Italy

pierpaolo.basile@uniba.it (P. Basile); pierluigi.cassotti@uniba.it (P. Cassotti); marco.polignano@uniba.it (M. Polignano); lucia.siciliani@uniba.it (L. Siciliani); giovanni.semeraro@uniba.it (G. Semeraro)

0000-0002-0545-1105 (P. Basile); 0000-0001-7824-7167 (P. Cassotti); 0000-0002-3939-0136 (M. Polignano); 0000-0002-1438-280X (L. Siciliani); 0000-0001-6883-1853 (G. Semeraro)

commitment to transparency and collaborative attitude towards the scientific community. We want to prove that it is possible to foster innovation and build effective LLMs using only open resources.

2. Language Adaptation Approaches

LLMs, such as GPT [3], Vicuna [4], LLaMA [5], or BLOOM [2], are trained on vast amounts of text data from diverse sources, which gives them a broad understanding of language and context. Nonetheless, it is important to note that the general knowledge inherent in these models might not be optimised for a particular language [6]. For this reason, Language Adaptation can strongly support the model's capacity to effectively navigate and address downstream tasks in a specific language. Language Adaptation of LLMs refers to the process of tuning a pre-trained LM to work effectively with a specific target language. In the scientific literature, different approaches for Language Adaptation have been recently proposed [7]. Among them, we can distinguish i) continuing the pre-training on new data [8], ii) creating a model adapter [9], iii) training a random subset of the model parameters [10].

In this work, we focus mainly on the MAD-X strategy [11] that has already been applied to BLOOM and proved to perform well in several languages as reported in [7]. Adapters were originally applied in NLP for parameter efficiency and quick fine-tuning of a base pre-trained Transformer model to new domains and tasks. [12] for the first time exploited adapters for transferring a pre-trained monolingual model to an unseen language by relying on learning new token-level embeddings. However, this solution does not scale to a large number of languages. Essentially, it is possible to adapt a pre-training multilingual LM to another unseen language L_u by 1) fine-tuning the model directly to a specific task in L_u; 2) using the obtained model for performing inference in L_u.

Due to the multilingual nature of the original model, it is not possible to obtain good performance in L_u since the model tends to balance many languages. On the contrary, the MAD-x strategy tries to tailor the original model to the target language by using an adapter. The idea is to fine-tune the original model using the masked language learning strategy instead of fine-tuning it on a specific task. This allows the use of unlabeled data written using the target language for fine-tuning the model to that specific language. The invertible adapters strategy provided by the MAD-X configuration facilitates the adaptation of BLOOM to the Italian language. The language adapter, located within each Transformer block, consists of a bottleneck adapter with down- and up-projection feedforward layers. Meanwhile, the invertible adapter works in the embedding layers to address the discrepancies between the vocabularies of the original and newly introduced languages.

3. Adaptation Pipeline

Starting from BLOOM, we build three models, i.e.:

- the BLOOM model with language adaptation for Italian and fine-tuned using the EVALITA training data and the instruction-based dataset Dolly (**B-it-D-E**);
- the BLOOM model with language adaptation fine-tuned using only the EVALITA training data (**B-it-E**);
- the BLOOM model without language adaptation fine-tuned with the EVALITA training data (**B-E**).

The adaptation and fine-tuning process is sketched in Figure 1.

To reduce computational costs, we decided to use the 1B version of the BLOOM model[1] (i.e., "BLOOM-1b7"). It has not been trained on any dialogue instruction as the counterpart BLOOMZ, and it does not contain the training data documents written in the Italian language. We follow the hypothesis that instruction-based fine-tuning should be performed after a phase of Language Adaptation, with instructions provided in the specific language of interest.

As a language adaptation strategy, we use MAD-X [7]. To produce a valuable model, we follow the suggestions of the authors of the paper, using default script parameters and selecting a sample of 100,000 sentences in Italian. We decided to sample data from the Filtered Oscar Dataset [13] for the Italian Language[2] released by [14].

Over the language-adapted models, we perform a general-purpose instruction-based fine-tuning step. Specifically, we use a version of the Dolly Instruction Dataset [15], which was adequately translated into Italian. For the translation, we opt to use an open-source tool[3] instead of a closed software. Dolly[4] is made of 15k high-quality human-generated prompt/response pairs specifically designed for instruction tuning LLMs. The dataset was authored by more than 5,000 Databricks[5] employees during March and April of 2023, and instructions are not copied from the web or other LLMs.

The instructions are mostly about Open/Closed Q&A, ExtractSummarize information, Brainstorming, Classification, and Creative writing. This fine-tuning step has

[1] https://huggingface.co/bigscience/bloom-1b7
[2] https://huggingface.co/datasets/gsarti/clean_mc4_it
[3] https://pypi.org/project/argostranslate/
[4] https://huggingface.co/datasets/databricks/databricks-dolly-15k
[5] https://www.databricks.com/

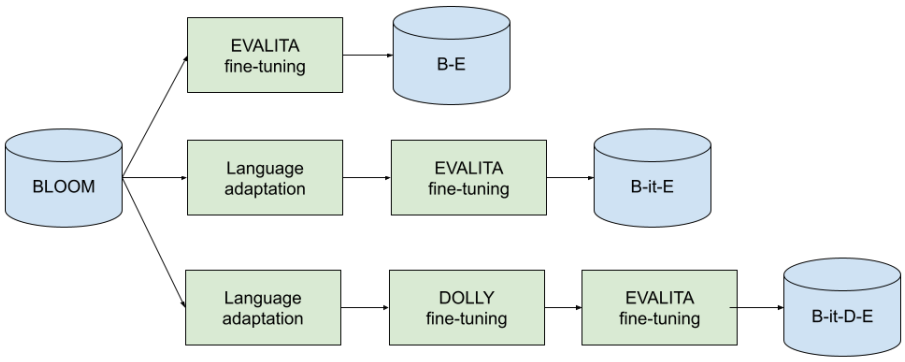

Figure 1: The adaptation pipeline

been performed by adapting the Python script released through GitHub [6].

Finally, we opt to fine-tune the models over two classification task prompts. To deal with this step, we decided to use data from two well-known EVALITA tasks, i.e., AMI2020 [16] and HaSpeeDe-v2-2020 [17]. AMI (Automatic Misogyny Identification) is aimed at automatically identifying misogyny in Tweets written in Italian. More specifically, Subtask A is focused on predicting Misogyny and Aggressiveness independently, while Subtask B is focused only on Misogyny but the dataset has been enriched with synthetic template-generated text. The HaSpeeDe task is focused on Hate Speech detection. The whole task is built on Tweets in Italian and Subtask A is aimed at determining whether the message contains Hate Speech or not, while Subtask B consists in determining whether the message contains Stereotype. All these tasks are structured as binary classification problems, where the label can be either `true` or `false`.

To use these resources to fine-tune our models, we first transformed the training data of the two tasks into an LLM prompt following a template. In particular, for the AMI task, we used the following template: *"instruction": "Nel testo seguente si esprime odio contro le donne? Rispondi sì o no.", "input": <training_text>, "output": <sì/no>*. Similarly, for HASPEEDE we used: *"instruction": "Il testo seguente incita all'odio? Rispondi sì o no.", "input": <training_text>, "output": <sì/no>*. To fill these templates, we mapped the label "1" with the word "sì" and the label "0" with the word "no", <training_text> is just the sentence from the dataset to classify. The fine-tuning step has been performed by using the same script as the previously described Dolly adaptation process.

[6] https://github.com/hyintell/BLOOM-fine-tuning/tree/main

3.1. Data Release

Following the open-science principles, we release all the models on HuggingFace. The available models are:

- **B-E**: the BLOOM model fine-tuned on EVALITA data without language adaptation[7];
- **B-it-E**: the BLOOM model adapted on the Italian language and fine-tuned on EVALITA data[8];
- **B-it-D-E**: the BLOOM model adapted on the Italian language and fine-tuned on both Dolly and EVALITA data[9].

It is important to underline that when you use the adapted LLM or one of its fine-tuned models is necessary to use the tokenizer of the adapted model. The BLOOM model adapted to the Italian language is available on HuggingFace[10].

The data used for fine-tuning are:

- the Italian translation of the Dolly dataset[11];
- the instructions generated from EVALITA data for training and test[12].

4. Validation and Discussion of Results

For the evaluation of the zero-shot abilities of the obtained models, we used the test data of AMI2020 and HASPEEDE-v2-2020. Also, in this case, the datasets have been translated into LLM prompts using the previously

[7] https://huggingface.co/basilepp19/bloom-1b7-evalita
[8] https://huggingface.co/basilepp19/bloom-1b7-it-evalita
[9] https://huggingface.co/basilepp19/bloom-1b7-it-dolly-evalita
[10] https://huggingface.co/basilepp19/bloom-1b7_it
[11] https://huggingface.co/datasets/basilepp19/dolly-15k-it
[12] https://huggingface.co/datasets/basilepp19/evalita2020-AH-instr

reported templates. The model output "sì/no" has been mapped back to the original labels 1/0 to allow a standard model evaluation setting.

Task	B-E	B-it-E	B-it-D-E	Baseline
AMI				
Subtask A	0.702	**0.730**	0.714	0.665
Subtask B	0.695	**0.785**	0.762	0.602
Haspeede2				
Task A (news)	0.518	0.555	**0.579**	0.621
Task A (tweets)	**0.706**	0.670	0.667	0.721
Task B (news)	0.584	0.623	**0.650**	0.669
Task B (tweets)	0.672	**0.686**	0.658	0.715

Table 1
Results

In Table 1, we report the results for AMI and Haspeede 2 tasks for all the three models: **B-E**, **B-it-E** and **B-it-D-E**. The evaluation metrics used are the Average Macro F1-score (F1) for AMI Subtask A and the Macro F1-score for Haspeede 2 tasks. The AMI Subtask A required predicting *Misogyny* and *Aggressiveness* classes independently using the Macro F1-score. The final score is obtained by averaging the two macro F1 scores. The results show that the MAD-X language adaptation with EVALITA fine-tuning **(B-it-E)** achieved the highest Average Macro F1-score of 0.730. The model that exploits Dolly fine-tuning obtains slightly worse results with an Average Macro F1-score of 0.714. In all the configurations, the model overcomes the task baseline. AMI Subtask B ranks model runs based on a weighted combination of AUC scores from the test raw dataset and three per-term AUC-based bias scores from the synthetic dataset, considering the performance for specific identity terms (e.g. "girlfriend" and "wife"). The results for Subtask B of the AMI task indicate that the BLOOM model with MAD-X language adaptation and EVALITA fine-tuning (**B-it-E**) achieved the highest Macro F1-score of 0.785, outperforming the other models and surpassing the task baseline significantly.

In Haspeede2 Task A, the models were required to classify *hateful* content, while Task B aimed to identify the presence of *stereotypes* related to the same targets, such as immigrants, Muslims and Roma. For the Haspeede2 Task A on news, BLOOM alone (**B-E**) achieved a moderate score of 0.522. The combination of MAD-X and EVALITA (**B-it-E**) improved the performance to 0.540, and adding Dolly fine-tuning (**B-it-D-E**) further increased it to 0.589. BLOOM alone yielded a reasonable score of 0.706 for tweets on Task A. While the other models which use language adaptation slightly decreased the performance.

In the Haspeede2 Task B evaluation, the MAD-X adaptation has proven to be remarkably effective for both news and tweets. Specifically, when applied to news data, it yielded outstanding results of 0.650 using both EVALITA and Dolly fine-tuning. Meanwhile, for tweets, the MAD-X adaptation achieved an even higher performance, reaching a score of 0.686 when using only the EVALITA fine-tuning. These findings highlight the adaptability and superiority of language adaptation (MAD-X) in handling different data types. All models cannot overcome the baselines in Haspeede2, but in three cases, the language adaptation provides the best result.

For fine-tuning and testing our models, we use a single NVIDIA A6000 GPU with 48 GB of RAM. The language adaptation steps require about 15 hours, while the fine-tuning of EVALITA 7 hours, and the Dolly fine-tuning only 5 hours.

5. Conclusions

In this paper, we explored a language adaptation strategy for the BLOOM model to address the challenge of handling languages not covered during the training.

Our approach is distinguished by its reliance on open-source data, software, and models, aligning with a commitment to transparency and accessibility in research. Despite the remarkable capabilities of the BLOOM model in understanding natural language for widely spoken languages, it showed limitations when applied to languages which are not included in the original training set, such as Italian. To overcome this limitation, we conducted experiments using the MAD-X language adaptation approach followed by instruction-based fine-tuning on Italian data.

The outcomes of our research demonstrate the effectiveness of language adaptation in significantly improving the zero-shot ability of the BLOOM model for Italian. The combination of MAD-X language adaptation with EVALITA fine-tuning achieved the highest performance on both the AMI2020 and HASPEEDE 2 tasks, showcasing the importance of the adaptation process for downstream classification tasks in Italian. In future work, we plan to evaluate our approach to more large BLOOM models and more recent tasks for the Italian.

The proposed methodology can be adapted for other languages and requires few examples for obtaining satisfying results. The adapted models can be easily fine-tuned on several tasks providing proper instructions. Our future research will extend to testing this approach with other languages and diverse adaptation strategies, contributing to the broader landscape of language model adaptability.

Acknowledgments

We acknowledge the support of the PNRR project FAIR - Future AI Research (PE00000013), Spoke 6 - Symbiotic AI (CUP H97G22000210007) under the NRRP MUR program funded by the NextGenerationEU.

References

[1] H. Naveed, A. U. Khan, S. Qiu, M. Saqib, S. Anwar, M. Usman, N. Barnes, A. Mian, A comprehensive overview of large language models, arXiv preprint arXiv:2307.06435 (2023).

[2] T. L. Scao, A. Fan, C. Akiki, E. Pavlick, S. Ilić, D. Hesslow, R. Castagné, A. S. Luccioni, F. Yvon, M. Gallé, et al., Bloom: A 176b-parameter open-access multilingual language model, arXiv preprint arXiv:2211.05100 (2022).

[3] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, Advances in neural information processing systems 33 (2020) 1877–1901.

[4] W.-L. Chiang, Z. Li, Z. Lin, Y. Sheng, Z. Wu, H. Zhang, L. Zheng, S. Zhuang, Y. Zhuang, J. E. Gonzalez, et al., Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, See https://vicuna. lmsys. org (accessed 14 April 2023) (2023).

[5] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al., Llama: Open and efficient foundation language models, arXiv preprint arXiv:2302.13971 (2023).

[6] K. Nowakowski, M. Ptaszynski, K. Murasaki, J. Nieuważny, Adapting multilingual speech representation model for a new, underresourced language through multilingual fine-tuning and continued pretraining, Information Processing & Management 60 (2023) 103148.

[7] Z.-X. Yong, H. Schoelkopf, N. Muennighoff, A. F. Aji, D. I. Adelani, K. Almubarak, M. S. Bari, L. Sutawika, J. Kasai, A. Baruwa, et al., Bloom+ 1: Adding language support to bloom for zero-shot prompting, arXiv preprint arXiv:2212.09535 (2022).

[8] E. C. Chau, L. H. Lin, N. A. Smith, Parsing with multilingual bert, a small corpus, and a small treebank, arXiv preprint arXiv:2009.14124 (2020).

[9] R. Wang, D. Tang, N. Duan, Z. Wei, X. Huang, G. Cao, D. Jiang, M. Zhou, et al., K-adapter: Infusing knowledge into pre-trained models with adapters, arXiv preprint arXiv:2002.01808 (2020).

[10] A. Ansell, E. M. Ponti, A. Korhonen, I. Vulić, Composable sparse fine-tuning for cross-lingual transfer, arXiv preprint arXiv:2110.07560 (2021).

[11] J. Pfeiffer, I. Vulić, I. Gurevych, S. Ruder, MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, 2020, pp. 7654–7673. URL: https://aclanthology.org/2020.emnlp-main.617. doi:10.18653/v1/2020.emnlp-main.617.

[12] M. Artetxe, S. Ruder, D. Yogatama, On the cross-lingual transferability of monolingual representations, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 4623–4637. URL: https://aclanthology.org/2020.acl-main.421. doi:10.18653/v1/2020.acl-main.421.

[13] J. Abadji, P. O. Suarez, L. Romary, B. Sagot, Towards a cleaner document-oriented multilingual crawled corpus, arXiv preprint arXiv:2201.06642 (2022).

[14] G. Sarti, M. Nissim, It5: Large-scale text-to-text pretraining for italian language understanding and generation, arXiv preprint arXiv:2203.03759 (2022).

[15] M. Conover, M. Hayes, A. Mathur, J. Xie, J. Wan, S. Shah, A. Ghodsi, P. Wendell, M. Zaharia, R. Xin, Free dolly: Introducing the world's first truly open instruction-tuned llm, 2023. URL: https://www.databricks.com/blog/2023/04/12/dolly-first-open-commercially-viable-instruction-tuned-llm.

[16] E. Fersini, D. Nozza, P. Rosso, et al., Ami@ evalita2020: Automatic misogyny identification, in: Proceedings of the 7th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2020), (seleziona...), 2020.

[17] M. Sanguinetti, G. Comandini, E. Di Nuovo, S. Frenda, M. Stranisci, C. Bosco, T. Caselli, V. Patti, I. Russo, Haspeede 2@ evalita2020: Overview of the evalita 2020 hate speech detection task, Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (2020).