Exploring Text-Embedding Retrieval Models for the Italian Language

Yuri Noviello[1], Fabio Tamburini[1]

[1] *FICLIT - University of Bologna, Via Zamboni, 32, Italy*

Abstract

Text retrieval systems have become essential in the field of natural language processing (NLP), serving as the backbone for applications such as search engines, document indexing, and information retrieval. With the rise of generative AI, particularly Retrieval-Augmented Generation (RAG) systems, the demand for robust text retrieval models has increased. However, existing large language models (LLMs) and datasets are often insufficiently optimized for Italian, limiting their performance in Italian text retrieval tasks. This paper addresses this gap by proposing both a data collection and specialized models tailored for Italian text retrieval. Through extensive experimentation, we analyze the improvements and limitations in retrieval performance, paving the way for more effective Italian NLP applications.

Keywords

Italian embedding, text embedding, retrieval model

1. Introduction

In recent years, text retrieval systems have emerged as a cornerstone of the natural language processing (NLP) field. These systems are crucial in various applications, including search engines, document indexing, and information retrieval tasks. Their primary function is to fetch relevant pieces of text from large corpora, enabling efficient and accurate information access. This capability is crucial for numerous industries, including legal, medical, and customer service sectors, where timely and precise information retrieval can significantly impact decision-making processes.

With the advent of generative AI, the importance of text retrieval systems has only amplified. Advanced systems, particularly chatbots based on Retrieval-Augmented Generation (RAG) [1], have become essential tools for various purposes. RAG systems combine retrieval mechanisms with generative models to produce contextually relevant and accurate responses in conversational AI applications. This integration has enhanced the capabilities of chatbots, making them more efficient in providing precise information and engaging in meaningful dialogues.

Despite the impressive performance of recent large language models (LLMs) as conversational agents in Italian contexts, there remains a notable gap in the resources and models specifically designed for Italian text retrieval tasks. This shortfall highlights a significant area for improvement and development within the Italian NLP community.

To address this gap, our work aims to propose both novel datasets and specialized models optimized for Italian text retrieval. By focusing exclusively on the Italian language, we strive to enhance the performance of retrieval tasks.

The primary contribution of this paper is the introduction of a comprehensive Italian text retrieval system, encompassing both a curated dataset collection and specialized language models. Through extensive experimentation and rigorous evaluation, we demonstrate the effectiveness of our approach, setting the stage for more advanced and reliable Italian text retrieval solutions applicable across diverse tasks.

2. Related Works

The development of text embedding models has seen significant advancements over the years, evolving from simple word representations to sophisticated contextual embeddings. Early models like Word2Vec [2] and GloVe [3] set the foundation by capturing semantic relationships between words through fixed-size vector representations. These models, however, lacked the ability to understand context, leading to the development of more advanced techniques.

Transformers have revolutionized the field of NLP by introducing mechanisms to capture context and relationships across entire sentences. BERT (Bidirectional Encoder Representations from Transformers [4]) marked a significant milestone, providing deep contextualized word embeddings by considering both left and right contexts simultaneously. This innovation has paved the way

CLiC-it 2024: Tenth Italian Conference on Computational Linguistics, Dec 04 — 06, 2024, Pisa, Italy

novielloyuri@gmail.com (Y. Noviello); fabio.tamburini@unibo.it (F. Tamburini)
https://www.unibo.it/sitoweb/fabio.tamburini (F. Tamburini)
0009-0006-5846-7756 (Y. Noviello); 0000-0001-7950-0347 (F. Tamburini)

for various large language models (LLMs), such as GPT-3 [5] and T5 [6], which further extend the capabilities of transformers by scaling up model size and training data.

Sentence Transformers, an extension of the transformer architecture [7], focus on generating embeddings for whole sentences rather than individual words. Models like SBERT (Sentence-BERT) enhance the performance of sentence-level tasks, such as semantic textual similarity and information retrieval, by fine-tuning BERT specifically for sentence embeddings. This approach has demonstrated significant improvements in capturing the semantic meaning of sentences, but specific training corpora, annotated with sentence similarity scores, must be provided for setting up the system.

In the realm of multilingual models, the multilingual E5 family has emerged as a robust solution for handling multiple languages within a single model architecture [8]. These models are pre-trained on a multilingual corpus, enabling them to perform effectively across different linguistic contexts. The multilingual E5 models leverage the strengths of transformer architectures to provide high-quality embeddings for numerous languages, including less-resourced ones. This makes them particularly valuable for tasks requiring cross-lingual understanding and retrieval.

The continuous evolution of text embedding models, from standard embeddings to advanced transformer-based approaches, highlights the dynamic nature of NLP research. Each progression addresses the limitations of its predecessors, contributing to more accurate and context-aware representations, which are crucial for a wide array of applications in natural language understanding and information retrieval.

3. Data

The quality and abundance of the data is one of the main aspect in order to obtain high quality text embedding models. The data used in this work for training the models were adapted from the following datasets: MIRACL [9], SQuAD-it [10], MLDR [11] and WikipediaQA-ita [12]. Among these, only the Multilingual Long-Document Retrieval (MLDR) was used as-is, as it already contains $2,151$ examples of Italian triplets in the form of `query-positive passage-negative passage`. Following sections detail the processing of the other datasets.

3.1. MIRACL-it

The Multilingual Information Retrieval Across a Continuum of Languages (MIRACL) dataset is widely used for building multilingual information retrieval models, such as the multilingual E5 models family [8]. Although the dataset encompasses 18 different languages, it does not include any Italian data. Given the dataset high quality, particularly in defining hard negatives through manual annotation, we decided to translate the dataset into Italian using automated methods. In particular, we focused on the English section of the dataset, which is organized as shown in Table 1.

Table 1
English data organization of MIRACL

Split	Query	Passage
train	2,863	29,416
dev	799	8,350
corpus	-	32,893,221

The translation process aimed to preserve these qualities while adapting the content to Italian, thereby creating a robust resource for training and evaluating Italian text retrieval models.

To translate the dataset, we experimented with two different approaches: a large language model (LLM) translation via the PaLM 2 API [13] and an open-source offline translation via Argos Translate [14]. The translation quality was evaluated to ensure that the Italian version maintained the dataset integrity and usefulness for training effective retrieval models.

3.1.1. Datasets translation using PaLM 2

We performed the translation of the whole training and development English sets of MIRACL using PaLM 2 API [13]. Due to budget constraints, we did not translate the entire corpus, as it would have required approximately €10,000, given the huge number of documents. We used the following prompt in order to obtain the Italian translation:

```
Translate the following text in Italian.
Write the translation only:
{text}
```

We used the same prompt for both queries and documents. For documents, we used the model `text-bison-32k@002`, and for queries, we relied on `text-bison@002`. This resulted in a total of $37,351$ API calls, as some documents are associated with multiple queries.

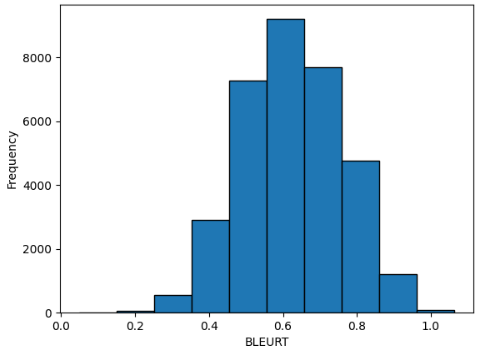
3.1.2. Open-source offline translation using Argos Translate

Argos Translate is an open-source library that uses OpenNMT for translation and supports multiple language model packages [14]. We utilized the English-to-Italian model to translate the training and development sets of MIRACL, including the entire corpus.

3.1.3. Translations quality evaluation

The translation performed by PaLM 2, as reported in the Technical Report [13] and confirmed by our empirical tests, is considered high-quality. To measure the quality of the translation performed by Argos Translate, we used the SOTA automatic metric BLEURT [15] and we used the PaLM 2 translations as reference. Since we do not have the entire corpus translated by the LLM, we conducted the evaluation only on the overlapping portion of the translated datasets, resulting in a corpus of $33,689$ documents.

Figure 1: BLEURT distribution



The average BLEURT score of 0.625 indicates that Argos Translate produced a decent translation, validating its use as a cost-effective alternative for text embedding model fine-tuning and evaluation.

3.2. SQuAD-it

SQuAD-it is obtained through semi-automatic translation of the SQuAD dataset into Italian, it contains more than $60,000$ question-answer pairs. For these experiments, we considered only the `question` and `context` attributes of each dataset example. Then, since we need triplets in the form of `query` - `positive passage` - `negative passage`, we performed hard negatives mining. We used the standard BM25 algorithm [16] to extract the top-10 similar documents for each query, excluding positive passages for the given query. This process ensured that the dataset was suitably challenging for training robust retrieval models.

3.3. WikipediaQA-ita

The WikipediaQA-ita is a datasets synthetically generated using a custom model from ReDiX Informatica; it has been created on Italian and specifically designed for RAG finetuning. It contains more than $100,000$ question-answer pairs. Similar to SQuAD-it, we considered only the `question` and `context` attributes for each example and applied the same hard negative mining strategy using the BM25 algorithm.

4. Methodology

4.1. Contrastive learning on labeled data

This work implements a dual-encoder model that uses a combination of supervised loss functions to achieve effective learning.

The dual-encoder model encodes queries and passages separately to produce their respective embeddings:

$$q_i = \text{Encoder}_{\text{query}}(Q_i) \tag{1}$$

$$p_j = \text{Encoder}_{\text{passage}}(P_j) \tag{2}$$

The similarity score between a query Q_i and a passage P_j is computed as the dot product of their embeddings:

$$S_{ij} = q_i \cdot p_j \tag{3}$$

The embeddings are normalized before computing the dot product, resulting in cosine similarity:

$$\hat{\mathbf{q}}_i = \frac{q_i}{\|q_i\|} \quad \text{and} \quad \hat{\mathbf{p}}_j = \frac{p_j}{\|p_j\|} \tag{4}$$

Thus, the similarity score becomes:

$$S_{ij} = \hat{\mathbf{q}}_i \cdot \hat{\mathbf{p}}_j \tag{5}$$

For a batch of queries and passages, the contrastive loss encourages higher similarity scores for matching query-passage pairs and lower scores for non-matching pairs. The loss function is defined as:

$$L_{\text{cont}} = \frac{1}{N} \sum_{i=1}^{N} \left[-\log \frac{\exp(S_{ii}/\tau)}{\sum_{j=1}^{N} \exp(S_{ij}/\tau)} \right] \tag{6}$$

where N is the batch size, τ is the temperature parameter, and S_{ii} represents the similarity score for the matching query-passage pair.

4.2. Fine-tuning procedure

We performed our answer-generation experiments by using the following base models:

1. `Minerva-1B` [17],
2. `Qwen2-1.5B` [18],
3. `Gemma-2B` [19],

We relied on the foundational versions of these models. To speed up the computation, we implemented a LoRA fine-tuning procedure. As a pooling strategy, we used EOS (End-Of-Sequence) pooling and normalized the embeddings. While we did not apply any prefix for passages, we added the following prefix to queries:

```
Given a search query, retrieve relevant
passages that answer the query.\nQuery:
```

We also experimented with using an Italian text prefix but found no significant difference in performance. Therefore, we opted for an English prefix to maintain consistency with other open-source models.

The fine-tuning process was executed on a weighted mixture of the datasets reported in Table 2. During this phase, the tokenization of the datasets documents was truncated at 512 tokens. We trained the model in mixed precision for 3 epochs, using a learning rate of 10^{-5}.

For each model, we conducted two fine-tuning experiments: one using the dataset with MIRACL data translated with PaLM 2 and another using the dataset translated with Argos Translate.

Table 2
Fine-tuning datasets organization

Source	Sample
MIRACL-it	100%
MLDR-it	100%
SQuAD-it	20%
WikipediaQA-ita	10%

4.3. Evaluation procedure

For the evaluation, we considered only the datasets for whose we already had the representation of relevance judgments (Qrels) in the TREC standard format [20], namely MIRACL-it and MLDR-it. This setup allows for a comprehensive evaluation of Retrieval Systems for the Italian language, encompassing both small/medium and large documents.

As with the training procedure, we evaluated each model using both the dataset with MIRACL data translated with PaLM 2 and the dataset translated with Argos Translate. To ensure consistency, we conducted evaluations only on the overlapping portions of the datasets between the two translations.

After creating the embeddings for both the test queries and documents, we used FAISS [21] to retrieve relevant documents. Finally, we employed the original implementation of TREC-eval for metrics computation.

We evaluated the models using the following metrics:

1. MRR@10 (Mean Reciprocal Rank): Measures the average of the reciprocal ranks of the first relevant document retrieved.
2. Recall@100: Measures the proportion of relevant documents retrieved among the top 100 results.
3. nDCG@10 (Normalized Discounted Cumulative Gain): Measures the ranking quality by comparing the order of results to the ideal ranking, emphasizing higher ranks.

5. Discussion and Analysis

We propose a comparison of the performance of different models on our Italian benchmark. For this analysis, we considered the Multilingual Sentence Transformers models [22] and the multilingual versions of the E5 models family. The scores are reported in Table 3.

5.1. Argos vs PaLM

By observing the performance on the MIRACL sets translated with PaLM 2 and Argos Translate, we found that every model achieved better results on the dataset translated with the PaLM 2 API. This behavior can be attributed to the higher translation quality provided by PaLM 2, which likely offers clearer sentence structures for the models to process.

However, since the difference in the results is very marginal, we can state that the machine translation provided by Argos Translate is a valid and cost-effective alternative for text embedding modeling.

On the contrary, we did not find any significant correlation between the models trained with different translation versions, given their small difference in scores, except for the MLDR-it evaluation of `gemma-2B-Argos`, which will be discussed later. This indicates that while translation quality can impact performance, the overall difference may not be substantial enough to render one method vastly superior to the other in practical applications for this specific task.

5.2. Multilingual Sentence Transformers

Generally, the performance of the Multilingual Sentence Transformers is similar when evaluated on the MIRACL-it sets. However, there is a notably significant performance gap for the MLDR-it dataset. We attribute the very poor performance of the `paraphrase-multi-MiniLM-L12-v2` model to its small maximum input token length of 128 tokens, which is unsuitable for datasets containing long documents. As expected, both our proposed models and the E5 models outperform all the Multilingual Sentence Transformers across all metrics on every dataset.

Table 3
Retrieval performance on Italian datasets

MODEL	MIRACL-it Argos 33k			MLDR-it test			MIRACL-it PaLM 33k		
	MRR	R	nDCG	MRR	R	nDCG	MRR	R	nDCG
distiluse-base-multi-cased-v1	56.83	92.32	52.47	16.44	58.50	18.74	58.20	93.22	54.05
distiluse-base-multi-cased-v2	51.20	88.34	46.73	15.52	54.00	17.48	51.68	90.06	47.83
paraphrase-multi-MiniLM-L12-v2	56.69	86.22	50.10	6.76	28.50	7.99	58.48	86.74	51.07
paraphrase-multi-mpnet-base-v2	62.26	93.59	57.22	15.14	50.00	17.70	63.02	94.00	58.03
multilingual-E5-base	*75.18*	*98.13*	*71.91*	40.24	66.00	42.55	75.66	*98.44*	*73.06*
multilingual-E5-large	**78.28**	**98.50**	**74.68**	40.56	71.50	43.38	**79.25**	**99.12**	**76.18**
minerva-1B-Argos	66.45	94.51	61.77	36.04	67.50	38.75	67.93	96.39	64.04
minerva-1B-PaLM	65.32	94.38	60.74	36.55	68.00	38.91	67.73	96.49	63.81
qwen2-1.5B-Argos	73.47	96.98	69.04	40.19	70.50	42.68	74.95	97.96	71.29
qwen2-1.5B-PaLM	73.16	97.21	69.12	**40.87**	69.00	**43.94**	74.54	98.04	70.56
gemma-2B-Argos	73.05	96.42	69.05	37.19	**75.00**	39.78	*75.80*	98.43	71.95
gemma-2B-PaLM	72.56	96.33	68.87	*40.75*	*74.50*	*43.46*	75.30	98.10	71.87

5.3. Multilingual E5 Models

The Multilingual E5 Models achieved very high scores in the evaluation of both datasets. In particular, the `multilingual-E5-large` model achieved the best MRR@10, Recall@100, and nDCG@10 scores on both translations of the MIRACL dataset. As expected, the `multilingual-E5-large` outperformed the base version, although the performance gap narrows with longer documents (MLDR-it).

5.4. Proposed Models

By observing the scores obtained by our proposed models, it appears that the models based on `Minerva-1B` achieved lower scores compared to the others, suggesting that it may not be the most suitable foundation model for this type of task.

The results obtained by the `Gemma-2B` and `Qwen2-1.5B` based models are very similar, except for the low MRR@10 and nDCG@10 scores obtained by `gemma-2B-Argos` on the MLDR-it dataset, which could indicate worse training stability caused by data translated with Argos Translate. However, the model achieved the best Recall@100 score on the same dataset, suggesting that this behavior may be caused by random noise during fine-tuning.

Finally, our proposed models achieved both the first and second best scores for each metric associated with the MLDR-it test set, demonstrating their effectiveness in handling long document retrieval tasks.

6. Conclusions

This work presents a comprehensive study on models and datasets focused on Information Retrieval (IR) for Italian documents. The primary contribution of this paper lies in illustrating a strategy for fine-tuning Large Language Models (LLMs) to achieve effective semantic representations of Italian texts. Additionally, we provide original models and datasets that serve as a starting point to bridge the performance gap between models designed for Italian and those optimized for other languages.

Our results demonstrate that the proposed models achieve performance comparable with state-of-the-art models for medium-sized documents and even surpass them when dealing with datasets containing very long documents. This suggests that our tailored approach to Italian text retrieval is not only viable but also highly effective.

6.1. Limitations and Future works

One of the main limitations of this study is the limited availability of hardware resources. Our fine-tuning process involved a significantly smaller number of dataset examples, well below $50,000$, compared to the multilingual E5 models, which were pre-trained on over 2 billion text pairs and fine-tuned on more than 1 million.

Additionally, we were unable to evaluate the proposed models on the complete MIRACL corpus, as it would have required more than 100 hours of computation per model. This restriction has highlighted a key area for potential improvement in our research. Future work could benefit significantly from experiments involving larger quantities of Italian data and the application of more advanced model architectures.

7. Online Resources

The fine-tuned adapters and the datasets have been made available (Models[1], Datasets[2]).

8. Implementation Details

All the experiments were executed on a Compute Engine Virtual Machine with 2 NVIDIA L4 GPUs.

8.1. Translation

While the offline translation relies on the model proposed by Argos Translated, to speed up computation, we directly utilized the API of CTranslate2 [23].

8.2. Fine-tuning

The fine-tuning experiments were conducted using an adaptation of the code from the Tevatron Toolkit [24]. The primary modifications included excluding the "title" attribute from document encoding to simulate a realistic scenario and filtering out queries not associated with negative passages.

8.3. Evaluation

Similar to the fine-tuning process, the evaluation was conducted without considering the "title" attribute for documents. Each model was evaluated according to the instructions provided by the authors. For creating embeddings with the Multilingual Sentence Transformers, we relied on the `sentence-transformers` implementation. For all other models, we used the `transformers` library [25].

Acknowledgments

We would like to thank Dinova Srl for funding this research and providing access to the Google Cloud Virtual Machines used in this project. Their support has been essential for this work.

Credit author statement

YN: Conceptualization, Investigation, Software, Formal analysis, Visualization, Writing - Original Draft.
FT: Methodology, Supervision, Writing - Review & Editing.

References

[1] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, D. Kiela, Retrieval-augmented generation for knowledge-intensive nlp tasks, in: Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS '20, Curran Associates Inc., Red Hook, NY, USA, 2020.

[2] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, Proceedings of Workshop at ICLR 2013 (2013).

[3] J. Pennington, R. Socher, C. Manning, GloVe: Global vectors for word representation, in: A. Moschitti, B. Pang, W. Daelemans (Eds.), Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Doha, Qatar, 2014, pp. 1532–1543. URL: https://aclanthology.org/D14-1162. doi:10.3115/v1/D14-1162.

[4] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: J. Burstein, C. Doran, T. Solorio (Eds.), Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: https://aclanthology.org/N19-1423. doi:10.18653/v1/N19-1423.

[5] 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/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

[6] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, Journal of Machine Learning Research 21 (2020) 1–67. URL: http://jmlr.org/papers/v21/20-074.html.

[7] N. Reimers, I. Gurevych, Sentence-BERT: Sentence embeddings using Siamese BERT-networks, in: K. Inui, J. Jiang, V. Ng, X. Wan (Eds.), Proceed-

[1] https://huggingface.co/collections/yuri-no/italian-retrieval-llm-adapters-667ab367ce13150b7c774078

[2] https://huggingface.co/collections/yuri-no/italian-retrieval-datasets-667acdccf922286634ef603b

ings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 3982–3992. URL: https://aclanthology.org/D19-1410. doi:10.18653/v1/D19-1410.

[8] L. Wang, N. Yang, X. Huang, L. Yang, R. Majumder, F. Wei, Multilingual e5 text embeddings: A technical report, 2024. arXiv:2402.05672.

[9] X. Zhang, N. Thakur, O. Ogundepo, E. Kamalloo, D. Alfonso-Hermelo, X. Li, Q. Liu, M. Rezagholizadeh, J. Lin, MIRACL: A Multilingual Retrieval Dataset Covering 18 Diverse Languages, Transactions of the Association for Computational Linguistics 11 (2023) 1114–1131. URL: https://doi.org/10.1162/tacl_a_00595. doi:10.1162/tacl_a_00595.

[10] D. Croce, A. Zelenanska, R. Basili, Neural learning for question answering in italian, in: C. Ghidini, B. Magnini, A. Passerini, P. Traverso (Eds.), AI*IA 2018 – Advances in Artificial Intelligence, Springer International Publishing, Cham, 2018, pp. 389–402.

[11] J. Chen, S. Xiao, P. Zhang, K. Luo, D. Lian, Z. Liu, Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation, 2024. arXiv:2402.03216.

[12] ReDiX-Informatica, wikipediaqa-ita: An open dataset of italian qa from wikipedia documents, https://https://huggingface.co/ReDiX/wikipediaQA-ita, 2024.

[13] R. Anil, A. M. Dai, O. Firat, M. Johnson, D. Lepikhin, A. Passos, S. Shakeri, E. Taropa, P. Bailey, Z. Chen, E. Chu, J. H. Clark, L. E. Shafey, Y. Huang, K. Meier-Hellstern, G. Mishra, E. Moreira, M. Omernick, K. Robinson, S. Ruder, Y. Tay, K. Xiao, Y. Xu, Y. Zhang, G. H. Abrego, J. Ahn, J. Austin, P. Barham, J. Botha, J. Bradbury, S. Brahma, K. Brooks, M. Catasta, Y. Cheng, C. Cherry, C. A. Choquette-Choo, A. Chowdhery, C. Crepy, S. Dave, M. Dehghani, S. Dev, J. Devlin, M. Díaz, N. Du, E. Dyer, V. Feinberg, F. Feng, V. Fienber, M. Freitag, X. Garcia, S. Gehrmann, L. Gonzalez, G. Gur-Ari, S. Hand, H. Hashemi, L. Hou, J. Howland, A. Hu, J. Hui, J. Hurwitz, M. Isard, A. Ittycheriah, M. Jagielski, W. Jia, K. Kenealy, M. Krikun, S. Kudugunta, C. Lan, K. Lee, B. Lee, E. Li, M. Li, W. Li, Y. Li, J. Li, H. Lim, H. Lin, Z. Liu, F. Liu, M. Maggioni, A. Mahendru, J. Maynez, V. Misra, M. Moussalem, Z. Nado, J. Nham, E. Ni, A. Nystrom, A. Parrish, M. Pellat, M. Polacek, A. Polozov, R. Pope, S. Qiao, E. Reif, B. Richter, P. Riley, A. C. Ros, A. Roy, B. Saeta, R. Samuel, R. Shelby, A. Slone, D. Smilkov, D. R. So, D. Sohn, S. Tokumine, D. Valter, V. Vasudevan, K. Vodrahalli, X. Wang, P. Wang, Z. Wang, T. Wang, J. Wieting, Y. Wu, K. Xu, Y. Xu, L. Xue, P. Yin, J. Yu, Q. Zhang, S. Zheng, C. Zheng, W. Zhou, D. Zhou, S. Petrov, Y. Wu, Palm 2 technical report, 2023. arXiv:2305.10403.

[14] P. Finlay, C. Argos Translate, Argos translate, 2021.

[15] T. Sellam, D. Das, A. Parikh, BLEURT: Learning robust metrics for text generation, 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. 7881–7892. URL: https://aclanthology.org/2020.acl-main.704. doi:10.18653/v1/2020.acl-main.704.

[16] S. Robertson, H. Zaragoza, The probabilistic relevance framework: Bm25 and beyond, Found. Trends Inf. Retr. 3 (2009) 333–389. URL: https://doi.org/10.1561/1500000019. doi:10.1561/1500000019.

[17] R. Orlando, L. Moroni, P.-L. H. Cabot, S. Conia, E. Barba, R. Navigli, Minerva llms: The first family of llms pretrained from scratch on italian., https://nlp.uniroma1.it/minerva/, 2024.

[18] Qwen2 technical report (2024).

[19] G. Team, T. Mesnard, C. Hardin, R. Dadashi, S. Bhupatiraju, S. Pathak, L. Sifre, M. Rivière, M. S. Kale, J. Love, P. Tafti, L. Hussenot, P. G. Sessa, A. Chowdhery, A. Roberts, A. Barua, A. Botev, A. Castro-Ros, A. Slone, A. Héliou, A. Tacchetti, A. Bulanova, A. Paterson, B. Tsai, B. Shahriari, C. L. Lan, C. A. Choquette-Choo, C. Crepy, D. Cer, D. Ippolito, D. Reid, E. Buchatskaya, E. Ni, E. Noland, G. Yan, G. Tucker, G.-C. Muraru, G. Rozhdestvenskiy, H. Michalewski, I. Tenney, I. Grishchenko, J. Austin, J. Keeling, J. Labanowski, J.-B. Lespiau, J. Stanway, J. Brennan, J. Chen, J. Ferret, J. Chiu, J. Mao-Jones, K. Lee, K. Yu, K. Millican, L. L. Sjoesund, L. Lee, L. Dixon, M. Reid, M. Mikuła, M. Wirth, M. Sharman, N. Chinaev, N. Thain, O. Bachem, O. Chang, O. Wahltinez, P. Bailey, P. Michel, P. Yotov, R. Chaabouni, R. Comanescu, R. Jana, R. Anil, R. McIlroy, R. Liu, R. Mullins, S. L. Smith, S. Borgeaud, S. Girgin, S. Douglas, S. Pandya, S. Shakeri, S. De, T. Klimenko, T. Hennigan, V. Feinberg, W. Stokowiec, Y. hui Chen, Z. Ahmed, Z. Gong, T. Warkentin, L. Peran, M. Giang, C. Farabet, O. Vinyals, J. Dean, K. Kavukcuoglu, D. Hassabis, Z. Ghahramani, D. Eck, J. Barral, F. Pereira, E. Collins, A. Joulin, N. Fiedel, E. Senter, A. Andreev, K. Kenealy, Gemma: Open models based on gemini research and technology, 2024. URL: https://arxiv.org/abs/2403.08295. arXiv:2403.08295.

[20] D. Harman, The text retrieval conferences (trecs), in: Proceedings of a Workshop on Held at Vienna, Virginia: May 6-8, 1996, TIPSTER '96, As-

sociation for Computational Linguistics, USA, 1996, p. 373–410. URL: https://doi.org/10.3115/1119018.1119070. doi:10.3115/1119018.1119070.

[21] M. Douze, A. Guzhva, C. Deng, J. Johnson, G. Szilvasy, P.-E. Mazaré, M. Lomeli, L. Hosseini, H. Jégou, The faiss library (2024). arXiv:2401.08281.

[22] Y. Yang, D. Cer, A. Ahmad, M. Guo, J. Law, N. Constant, G. Hernandez Abrego, S. Yuan, C. Tar, Y.-h. Sung, B. Strope, R. Kurzweil, Multilingual universal sentence encoder for semantic retrieval, in: A. Celikyilmaz, T.-H. Wen (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, Association for Computational Linguistics, Online, 2020, pp. 87–94. URL: https://aclanthology.org/2020.acl-demos.12. doi:10.18653/v1/2020.acl-demos.12.

[23] OpenNMT, Ctranslate2, https://github.com/OpenNMT/CTranslate2, 2019.

[24] L. Gao, X. Ma, J. Lin, J. Callan, Tevatron: An efficient and flexible toolkit for neural retrieval, in: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '23, Association for Computing Machinery, New York, NY, USA, 2023, p. 3120–3124. URL: https://doi.org/10.1145/3539618.3591805. doi:10.1145/3539618.3591805.

[25] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, A. M. Rush, Transformers: State-of-the-Art Natural Language Processing, Association for Computational Linguistics, 2020, pp. 38–45. URL: https://www.aclweb.org/anthology/2020.emnlp-demos.6.