IT5: Large-scale Text-to-text Pretraining for Italian Language Understanding and Generation

Sarti, Gabriele; Nissim, Malvina

Published in:
ArXiv

Publication date:
2022

IT5: Large-scale Text-to-text Pretraining for Italian Language Understanding and Generation

Gabriele Sarti and **Malvina Nissim**
Center for Language and Cognition (CLCG)
University of Groningen
{g.sarti, m.nissim}@rug.nl

Abstract

The T5 model and its unified text-to-text paradigm contributed in advancing the state-of-the-art for many natural language processing tasks. While some multilingual variants of the T5 model have recently been introduced, their performances were found to provide suboptimal performances for languages other than English if compared to monolingual variants. We are motivated by these findings to introduce IT5, the first family of encoder-decoder transformer models pretrained specifically on Italian. We perform a thorough cleaning of a web-crawled Italian corpus including more than 40 billion words and use it to pretrain three IT5 models of different sizes. The performance of IT5 models and their multilingual counterparts is then evaluated on a broad range of natural language understanding and generation benchmarks for Italian. We find the monolingual IT5 models to provide the best scale-to-performance ratio across tested models, consistently outperforming their multilingual counterparts and setting a new state-of-the-art for most Italian conditional language generation tasks. Code, data and checkpoints are made available: https://github.com/gsarti/it5

1 Introduction

The text-to-text paradigm introduced by the T5 model (Raffel et al., 2020) has recently been widely adopted as a simple yet powerful generic transfer learning approach for most language processing tasks (Sanh et al., 2021; Aribandi et al., 2022). Although the original T5 model was trained exclusively on English data, the same architecture has been extended to a massively multilingual setting covering more than 100 languages by mT5 and ByT5 (Xue et al., 2021a,b), following recent advances in the multilingual pre-training of large language models such as mBERT, XLM, XLM-R and mDeBERTa (Devlin et al., 2019; Conneau and Lample, 2019; Conneau et al., 2020; He et al., 2021). Multilingual language models were shown to excel in cross-lingual and low-resource scenarios, but multiple studies have highlighted their suboptimal scale-to-performance ratio if compared to monolingual counterparts for language-specific applications in which data are abundant (see Nozza et al. 2020 and Rust et al. 2021 for an overview).

Following this line of research, Nagoudi et al. (2021) pre-trained several T5 models on the Arabic language, obtaining an improvement in performances from the multilingual mT5 model (and other transformer variants) on Arabic conditional language generation tasks such as news summarization, headline generation and question generation. Similarly, Carmo et al. (2020) pre-trained a Portuguese T5 model and showed competitive performances on named entity recognition and natural language inference.

In this work, we follow an approach similar to the one of Nagoudi et al. (2021) to pre-train and evaluate a set of three Italian T5 models of different sizes, which we identify under the name IT5. In Section 3 we present the procedure we adopted to clean the Italian mC4 corpus (Xue et al., 2021b) used for pre-training the IT5 models, and we briefly cover the set of pre-training parameters. Section 4 describes the multilingual baselines and the downstream tasks we used to evaluate the performances of fine-tuned IT5 models and presents the obtained results. Finally, we discuss our findings and future directions in Section 5. To the best of our knowledge, the IT5 models are the first publicly available encoder-decoder models pre-trained exclusively on the Italian language, and we view them as a substantial contribution towards the advancement of Italian NLP. Overall, our contributions are the following:

- We introduce a thoroughly cleaned version of the Italian mC4 corpus and use it pre-train three IT5 models of various dimensions.

- We evaluate the IT5 models on a broad set of natural language understanding and generation tasks for Italian, showing consistent improvements compared to multilingual models and the current state-of-the-art.

- We publicly release all the code, data, pre-trained and fine-tuned checkpoints for further experimentation by the research community.

2 Background

2.1 Text-to-text Transfer Transformers

The Text-to-text Transfer Transformer (T5) model introduced by Raffel et al. (2020) adapts the original Transformer architecture proposed by Vaswani et al. (2017) to a scenario where multiple natural language processing tasks are first reformulated into a unified text-to-text format and then used alongside masked span prediction for semi-supervised pre-training. The encoder-decoder architecture of T5 is especially suited for sequence-to-sequence tasks (Sutskever et al., 2014), which cannot be performed by encoder-only models like BERT (Devlin et al., 2019) and can prove to be challenging for decoder-only models like GPTs (Radford et al., 2019; Brown et al., 2020) due to the lack of explicit conditioning on source context. The same architecture can be easily extended to most natural language understanding tasks by converting them to the text-to-text format, making the T5 model highly versatile to most NLP settings.

2.2 Pre-trained Language Models for Italian

The rapid progress achieved in recent years by the NLP research community on shared benchmarks such as SuperGLUE (Wang et al., 2019) is primarily driven by model scaling and the large-scale pre-training, two components requiring both high technical expertise and heavy computational resources. These strict conditions significantly contributed to exacerbating inequalities in the availability of state-of-the-art systems across languages other than English. This is especially true in the case of the Italian language, which despite not being "low-resource" under any possible definition of the term, can currently count on a minimal set of publicly available pre-trained language models (Miaschi et al., 2021), and most notably on a single pre-trained model for text generation (GePpeTto by De Mattei et al. 2020a, built upon the GPT-2 small model architecture).

3 Data and Model Pretraining

The original T5 model is pre-trained on the Colossal Clean Crawled Corpus (C4) (Raffel et al., 2020), a large collection including roughly 750GB of web-scraped English texts sourced from the Common-Crawl. This was subsequently cleaned employing heuristics aimed at removing templated fillers, text deduplication, Javascript code, slurs and non-English texts. The multilingual counterpart of T5 adopts a similar procedure to create mC4 (Xue et al., 2021b), a multilingual version of the C4 corpus including 107 languages. While the authors of mC4 still perform text deduplication, language detection and bad words removal, they lower the language detection threshold to 70% and omit other useful heuristics due to the great variability across the covered character systems, such as filtering sentences having non-standard end-of-sentence punctuation. As a consequence, the resulting corpus has an overall lower quality, with a recent study finding 16% of examples in a random sample of mC4 being associated to the wrong language tag, and 11% of them not containing any linguistic information (Kreutzer et al., 2022).

In light of this, we decided to perform a more thorough data cleaning of the Italian portion of the mC4 before pre-training our IT5 models, as described in the following section.

3.1 Cleaning the Italian mC4 Corpus

The original Italian portion of the mC4 Corpus includes 1024 training files in JSON format of roughly 220MB each and eight validation files of roughly 24MB each, for a total of approximately 215GB of raw text, making it the most extensive publicly available Italian corpus to this date. We use a public implementation[1] reproducing and augmenting the original C4 data cleaning pipeline to perform sentence tokenization inside every document and remove sentences containing:

- words from a manually-selected subset of the Italian and English List of Dirty Naughty Obscene and Otherwise Bad Words;[2]

- less than three words, or a word longer than 1000 characters;

- an end symbol not matching standard end-of-sentence punctuation for Italian;

[1] https://gitlab.com/yhavinga/c4nlpreproc
[2] https://github.com/LDNOOBW

- strings associated to Javascript code, lorem ipsum, English and Italian privacy policy/cookie disclaimers.

After filtering sentences and recomposing documents from the remaining sentences, we further exclude documents containing less than five sentences or containing either less than 500 or more than 50,000 characters or not being identified as prevalently Italian by the LangDetect library.[3] The cleaning process was performed by using a parallel processing pipeline on a TPUv3-8 virtual machine with 96 CPU cores on Google Cloud Platform and required roughly 10 hours of computation due to the demanding steps of sentence tokenization and language detection. The size of the resulting corpus, which we name Clean Italian mC4 Corpus[4], is roughly halved by this cleaning procedure and includes approximately 215GB of raw text in Italian. This corresponds to a rough estimate of 103 million documents or 41 billion words.

3.2 Model and Training Parameters

Having obtained a clean corpus for pre-training, we train a SentencePiece unigram subword tokenizer (Kudo, 2018) with a vocabulary size of 32'000 words on the first 10 million cleaned documents and we use it to pre-train three models following the canonical small, base and large sizes defined in Raffel et al. (2020). Table 1 provides a brief overview of size-specific parameters, including training times and the number of steps, while Appendix B provides more comprehensive information about the overall architecture. Models are trained on a TPU v3-8 accelerator on Google Cloud Platform using the JAX framework (Bradbury et al., 2018) and Huggingface Transformers (Wolf et al., 2020). We adopt the T5 v1.1 architecture[5] also used by the mT5 model, improving upon the original T5 by using GeGLU nonlinearities (Shazeer, 2020), scaling model hidden size alongside feed-forward layers and pre-training only on unlabeled data, without dropout. All models are pre-trained with a learning rate of 5e-3 and a maximum sequence length of 512 tokens using the Adafactor optimizer (Shazeer and Stern, 2018) to reduce the

	IT5 Small	IT5 Base	IT5 Large
# of parameters	60M	220M	738M
# of steps	1'050'000	1'050'000	2'100'000
Training time	36 h	101 h	370 h
Batch size	128	128	64
Weight decay	1e-3	1e-3	1e-2

Table 1: Size-specific parameters for the three pre-trained IT5 models.

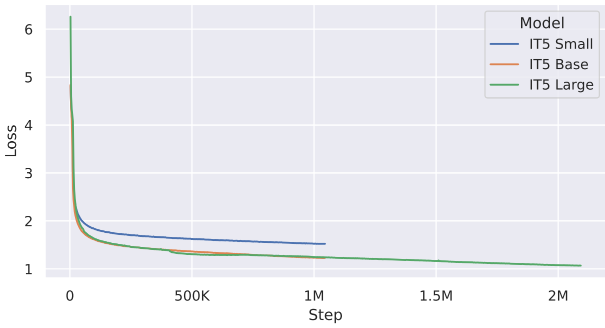

Figure 1: Loss curves for the masked span prediction task used to pre-train the IT5 models.

memory footprint of training and are validated on a fixed subset of 15'000 examples. Figure 1 shows the computed loss during the training process for the three models. We used the Google Cloud Carbon Footprint tool to estimate the overall amount of CO2 generated by the pre-training process and found it to be approximately equal to 7kgCO2, corresponding approximately to the emissions of a 60km car ride.[6]

4 Experiments

4.1 Fine-tuning Tasks

The following paragraphs provide an overview of the seven tasks we selected to evaluate the downstream performances of fine-tuned IT5 and mT5 models. Given the limited availability of evaluation resources for natural language generation in Italian, our selection aims to provide a comprehensive overview of canonical areas of the conditional text generation paradigm such as summarization, style transfer and question generation. Moreover, we also include a question answering task to provide a direct comparison of IT5 performances against encoder-based extractive systems.

Wikipedia Summarization To evaluate summarization capabilities in the encyclopedic domain,

[3] https://github.com/Mimino666/langdetect

[4] https://huggingface.co/datasets/gsarti/clean_mc4_it

[5] https://github.com/google-research/text-to-text-transfer-transformer/blob/main/released_checkpoints.md#t511

[6] https://ec.europa.eu/eurostat/cache/metadata/en/sdg_12_30_esmsip2.htm

we use the Wikipedia for Italian Text Summarization (WITS) corpus (Casola and Lavelli, 2022), containing roughly 700'000 articles extracted from a cleaned dump of the Italian Wikipedia alongside their leading sections, that are treated as summaries for the summarization task. We mimick the same evaluation setup as the original authors by using a test set of 10'000 examples.

News Summarization We evaluate the news article summarization setting by concatenating two collections of articles from the Fanpage.it and Il-Post newspapers scraped by the Applied Recognition Technology Laboratory[7] and published on the Huggingface Datasets Hub (Lhoest et al., 2021) alongside monolingual and multilingual modeling baselines. We refer to this concatenated corpus as NewsSum-IT from here onwards. We fine-tune our systems on the whole training set, including more than 100'000 articles and respective short summaries, and evaluate them separately on the two test sets defined by the dataset creators. We report the averaged metrics across the two newspapers in the results section.

Question Answering We evaluate our models on extractive question answering using the SQuAD-IT dataset (Croce et al., 2018), which consists of a subset of roughly 50'000 paragraph-question-answers triplets produced by automatically translating the original SQuAD dataset (Rajpurkar et al., 2016) and filtering out problematic examples. For using the dataset with our models and baselines, we frame the QA task as a sequence-to-sequence problem aimed at generating plausible responses given a source text in the format `<PARAGRAPH> Domanda: <QUESTION>`. We use the canonical splits for training and testing and evaluate generated answers using the script provided by the original authors.

Question Generation We leverage the same dataset used for QA to evaluate question generation capabilities by reversing the text triplets, making the model predict a plausible question given a source text in the format `<PARAGRAPH> Risposta: <ANSWER>`, where the answer is the first among the available answers for the given example. We adopt the same splits used for QA.

News Headline Style Transfer We evaluate style transfer abilities in the news domain on the CHANGE-IT shared task (De Mattei et al., 2020b), which includes two sets of roughly 60'000 newspaper articles and headlines each extracted from the left-leaning Italian newspaper la Repubblica and the right-leaning Il Giornale, respectively. We train and validate our models on the canonical splits using the same cross-source article-to-headline generation approach used for the original baselines by the authors since we find it to work more effectively in the absence of high-quality article alignments across all examples of the two corpora. We report the averaged performances for the two style transfer directions (Il Giornale ↔ la Repubblica).

News Headline Generation We combine the two CHANGE-IT subsets to create a corpus of roughly 120'000 news articles and headlines pairs, which we refer to with the name HeadGen-IT. We use HeadGen-IT to evaluate news headline generation performances of our models, using the same test splits defined in CHANGE-IT for prediction.

Formality Style Transfer We evaluate the formality style transfer capabilities of our models on the Italian subset of the XFORMAL dataset (Briakou et al., 2021), containing a training set of 115'000 forum messages automatically translated from the GYAFC corpus (Rao and Tetreault, 2018) and covering the topics of entertainment, music, family and relationships, and a small test set of 1000 formal-informal pairs obtained directly in Italian from four crowdworkers via Amazon Mechanical Turk. We evaluate our models in both style transfer directions (Formal ↔ Informal).

4.2 Evaluation Metrics

We adopt a common combination of text-based and trained metrics as a default for evaluating model performances in most settings. We use the language-independent ROUGE metric (Lin, 2004) in its unigram (R1), bigram (R2) and Longest Common Subsequence (RL) variants to evaluate lexical matches, and the trained BERTScore metric (Zhang et al. 2020; BS) to evaluate correspondence at the semantic level. Since BERTScore requires a pre-trained BERT-like model for the target language, we use a popular BERT model pre-trained on Italian for the evaluation. Following the authors' recommendation[11], we compute and use baseline scores (see Appendix C for details) in order to

[7] https://huggingface.co/ARTeLab

[11] github.com/Tiiiger/bert_score/blob/master/journal/rescale_baseline.md

	Size	WITS			
	#	R1	R2	RL	BS
TextRank (2022)	-	.302	.076	.197	-
LexRank (2022)	-	.269	.059	.175	-
SumBasic (2022)	-	.206	.048	.140	-
IT5 Small (2022)	60M	.216	.097	.193	-
mT5 Small (ours)	300M	.347	.200	.316	.517
mT5 Base (ours)	580M	.348	.200	.315	.520
IT5 Small (ours)	60M	.337	.191	.306	.504
IT5 Base (ours)	220M	**.369**	**.217**	**.333**	**.530**
IT5 Large (ours)	738M	.335	.191	.301	.508

	Size	SQuAD-IT QA	
	#	F1	EM
DrQA-IT (Croce et al., 2018)	-	.659	.561
mBERT (Croce et al., 2019)	110M	.760	.650
BERT [8] (Devlin et al., 2019)	110M	.753	.638
MiniLM (Riabi et al., 2021)	66M	.720	.577
$\text{MiniLM}_{+\texttt{st}}$ (2021)	66M	.745	.620
$\text{XLM-R Large}_{+\texttt{st}}$ (2021)	560M	**.804**	.676
mT5 Small (ours)	300M	.660	.560
mT5 Base (ours)	580M	.757	.663
IT5 Small (ours)	60M	.716	.619
IT5 Base (ours)	220M	.761	.663
IT5 Large (ours)	738M	.780	**.691**

	Size	NewsSum-IT			
	#	R1	R2	RL	BS
mBART Large [9,10]	610M	**.377**	**.194**	**.291**	-
mT5 Small (ours)	300M	.323	.150	.248	.375
mT5 Base (ours)	580M	.340	.161	.262	.393
IT5 Small (ours)	60M	.330	.155	.258	.386
IT5 Base (ours)	220M	.339	.160	.263	-
IT5 Large (ours)	738M	.251	.101	.195	.315

	Size	SQuAD-IT QG			
	#	R1	R2	RL	BS
mT5 Small (ours)	300M	.306	.143	.286	.463
mT5 Base (ours)	580M	.346	.174	.324	.495
IT5 Small (ours)	60M	.367	.189	.344	.505
IT5 Base (ours)	220M	.382	.199	.354	.516
IT5 Large (ours)	738M	**.383**	**.204**	**.360**	**.522**

Table 2: Performance of IT5 models, mT5 models and baselines on Wikipedia summarization (WITS), news summarization (NewsSum-IT), question answering (SQuAD-IT QA) and question generation (SQuAD-IT QG) for Italian. Best scores are marked in **bold**.

broaden the range of the metric and remove noise. For the QA task, we instead use the exact-match (EM) and F1-score (F1) metrics routinely used for evaluation. Finally, for the news headline style transfer task, we use trained classifiers provided by the authors[12] to ensure headline-headline (HH) and headline-article (HA) coherence.

4.3 Baselines

For every selected task, we compare the performances of the three IT5 models to those previously reported in the literature. Moreover, we adopt the same fine-tuning procedure for training and evaluating two sizes (small and base) of the multilingual T5 model (mT5) (Xue et al., 2021b). This allows us to assess the validity of our pre-training procedure and to observe whether the monolingual setting yields a performance improvement. We emphasize that mT5 models are considerably larger than regular T5 architectures due to a larger vocabulary and embedding matrix. While this fact should not affect their computational efficiency, their size in memory could limit their utilization by most NLP practitioners. Additional details on the parametrization across fine-tuning tasks for mT5 and IT5 models are provided in Appendix D.

4.4 Results and Discussion

Tables 2, 3 and 4 present the results of our fine-tuning experiments. Given the broad scope of our analysis, we limit ourselves to comment salient trends we observe across tasks.

IT5 models provide state-of-the-art performances for language generation and understanding tasks in Italian. The IT5 models outperform multilingual models and previous systems in 6 out of 8 evaluated tasks, with noticeable improvements over mT5 systems in particular for question answering and generation and for headline-headline coherence on the news headline style transfer task. More specifically to the QA setting, the IT5 Large model establishes a new top score for exact matches, outperforming most extrac-

[12] `https://github.com/michelecafagna26/CHANGE-IT`

[8] `https://huggingface.co/antoniocappiello/bert-base-italian-uncased-squad-it`

[9] `https://huggingface.co/ARTeLab/mbart-summarization-ilpost`

[10] `https://huggingface.co/ARTeLab/mbart-summarization-fanpage`

	Params	XFORMAL (IT) F → I				XFORMAL (IT) I → F				HeadGen-IT			
	#	R1	R2	RL	BS	R1	R2	RL	BS	R1	R2	RL	BS
mT5 Small (ours)	300M	.651	**.450**	.631	.666	.638	.446	.620	.684	.277	.094	.244	.408
mT5 Base (ours)	580M	**.653**	.449	**.632**	**.667**	.661	.471	.642	.712	.302	.109	.265	.427
IT5 Small (ours)	60M	.650	**.450**	.631	.663	.646	.451	.628	.702	.287	.100	.253	.414
IT5 Base (ours)	220M	.652	.446	.632	.665	.583	.403	.561	.641	**.310**	.112	**.270**	**.433**
IT5 Large (ours)	738M	.611	.409	.586	.613	**.663**	**.477**	**.645**	**.714**	.308	**.113**	**.270**	.430

Table 3: Performance of IT5 models and mT5 models on Italian formality style transfer (XFORMAL IT) for the formal-to-informal (F → I) and informal-to-formal (I → F) directions and on the news headline generation task (HeadGen-IT). Best scores are marked in **bold**.

	CHANGE-IT			
	HH	HA	RL	BS
PointerNet (2020b)	.644	.874	.151	-
BiLSTM$_{+\texttt{Att}}$ (2020c)	.744	.846	.155	-
mT5 Small (ours)	.777	.807	.211	.372
mT5 Base (ours)	.795	.799	.236	.398
IT5 Small (ours)	.898	**.882**	.231	.392
IT5 Base (ours)	**.904**	.868	**.247**	**.411**
IT5 Large (ours)	.895	.861	.237	.390

Table 4: Performance of IT5 models, mT5 models and baselines on news headline style transfer (CHANGE-IT) in Italian. Best scores are marked in **bold**.

tive systems and closing the gap with the XLM-R Large system adopted by Riabi et al. (2021), although the latter is trained on more data using a specific synthetic data augmentation procedure.

Multilingual models can still be helpful in specific applications and when using translated data. We observe that multilingual language models are still obtaining the best performances in the news summarization and the formal-to-informal style transfer tasks. In the case of news summarization, we attribute the performance gap in large part to the scale of the mBART baseline model. For the formality style transfer task, instead, after a minimal error analysis, we conjecture that translation errors and English acronyms present in the noisy training split of XFORMAL act as out-of-distribution samples in the monolingual setting, disrupting the performances of IT5 systems, but are captured more easily by multilingual systems which were exposed by multiple data distributions by design. This would indicate a better fit of multilingual pre-trained models for such settings if verified. We leave a more thorough analysis assessing the systematicity of these patterns to future work.

Scaling model size does not guarantee an increase in performance if not supported by an increase in computational resources. Contrary to the findings of "scaling laws" (Brown et al., 2020) related to model size for pre-trained Transformer models, we do not observe a systematic increase in downstream performances for base and large IT5 models across all tasks when compared to their small counterparts, despite reaching lower loss values and higher pre-training accuracies. While recent work highlighted how better pre-training performances do not always correspond to better downstream scores for T5 models (Tay et al., 2022), we hypothesize that in our case, the lack of improvement is related mainly to insufficient computational resources that limited the maximal batch size during training to 128, as opposed to the 2048 figure reported in the original work by Raffel et al. (2020).

5 Conclusion

This paper introduced IT5, the first family of large-scale encoder-decoder models pre-trained on the Italian language. We presented a detailed overview of the pre-training procedure, including the previous data cleaning steps, and performed a thorough evaluation of the IT5 models and their multilingual mT5 counterparts on a broad set of conditional language generation tasks for the Italian language. We obtained new state-of-the-art results across most evaluated tasks and concluded by discussing the shortcomings of large-scale monolingual language modeling when dealing with automatically translated data and limited computational resources.

In light of our results, we find the study of time and computation trade-offs between the pre-training of monolingual models and a language-specific continued pre-training of multilingual models promising to close the language gap in the performances of NLP models.

6 Broader Impact Statement

Despite our thorough cleaning procedure aimed at removing vulgarity and profanity, it must be acknowledged that models trained on web-scraped contents such as IT5 will inevitably reflect and amplify biases present in Internet blog articles and comments, resulting in potentially harmful content such as racial or gender stereotypes and conspiracist views. In light of this, the study of such biases is explicitly encouraged, and model usage should ideally be restricted to research-oriented and non-user-facing endeavors.

7 Acknowledgments

We thank the Google TPU Research Cloud program for providing us with free access to TPU v3-8 machines used in pre-training the IT5 models and the Center for Information Technology of the University of Groningen for providing access to the Peregrine high performance computing cluster used in fine-tuning and evaluation experiments. We also thank the Huggingface team for creating the original template for TPU-compatible pre-training and fine-tuning scripts for T5 models and making them available during the Huggingface JAX/Flax Community Week.

References

Vamsi Aribandi, Yi Tay, Tal Schuster, Jinfeng Rao, Huaixiu Steven Zheng, Sanket Vaibhav Mehta, Honglei Zhuang, Vinh Q. Tran, Dara Bahri, Jianmo Ni, Jai Gupta, Kai Hui, Sebastian Ruder, and Donald Metzler. 2022. ExT5: Towards extreme multi-task scaling for transfer learning. In *Proceedings of the Tenth International Conference on Learning Representations (ICLR'22)*.

James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. 2018. JAX: composable transformations of Python+NumPy programs.

Eleftheria Briakou, Di Lu, Ke Zhang, and Joel Tetreault. 2021. Olá, bonjour, salve! XFORMAL: A benchmark for multilingual formality style transfer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3199–3216, Online. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Diedre Carmo, Marcos Piau, Israel Campiotti, Rodrigo Nogueira, and Roberto de Alencar Lotufo. 2020. PTT5: Pretraining and validating the T5 model on brazilian portuguese data. *ArXiv Computation and Language*, arXiv:2008.09144(v2).

Silvia Casola and Alberto Lavelli. 2022. WITS: Wikipedia for italian text summarization. In *Proceedings of the Eight Italian Conference on Computational Linguistics (CLiC-it 2021)*, Milan, Italy. CEUR.org.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Danilo Croce, Giorgio Brandi, and Roberto Basili. 2019. Deep bidirectional transformers for italian question answering. In *Proceedings of the Sixth Italian Conference on Computational Linguistics (CLiC-it 2019)*, Bari, Italy. CEUR.org.

Danilo Croce, Alexandra Zelenanska, and Roberto Basili. 2018. Neural learning for question answering in italian. In *AI*IA 2018 – Advances in Artificial Intelligence*, pages 389–402, Cham. Springer International Publishing.

Lorenzo De Mattei, Michele Cafagna, Felice Dell'Orletta, Malvina Nissim, and Marco Guerini. 2020a. GePpeTto carves italian into a language model. In *Proceedings of the Seventh Italian Conference on Computational Linguistics (CLiC-it 2020)*, Online. CEUR.org.

Lorenzo De Mattei, Michele Cafagna, Felice Dell'Orletta, Malvina Nissim, and Albert Gatt. 2020b. CHANGE-IT @ EVALITA 2020: Change headlines, adapt news, generate. In *Proceedings of Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020)*, Online. CEUR.org.

Lorenzo De Mattei, Michele Cafagna, Huiyuan Lai, Felice Dell'Orletta, Malvina Nissim, and Albert Gatt. 2020c. On the interaction of automatic evaluation and task framing in headline style transfer. In *Proceedings of the 1st Workshop on Evaluating NLG Evaluation*, pages 38–43, Online (Dublin, Ireland). Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. *ArXiv Computation and Language*, arXiv:2111.09543(v2).

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Ortiz Suarez, Iroro Orife, Kelechi Ogueji, Andre Niyongabo Rubungo, Toan Q. Nguyen, Mathias Müller, André Müller, Shamsuddeen Hassan Muhammad, Nanda Muhammad, Ayanda Mnyakeni, Jamshidbek Mirzakhalov, Tapiwanashe Matangira, Colin Leong, Nze Lawson, Sneha Kudugunta, Yacine Jernite, Mathias Jenny, Orhan Firat, Bonaventure F. P. Dossou, Sakhile Dlamini, Nisansa de Silva, Sakine Çabuk Ballı, Stella Biderman, Alessia Battisti, Ahmed Baruwa, Ankur Bapna, Pallavi Baljekar, Israel Abebe Azime, Ayodele Awokoya, Duygu Ataman, Orevaoghene Ahia, Oghenefego Ahia, Sweta Agrawal, and Mofetoluwa Adeyemi. 2022. Quality at a Glance: An Audit of Web-Crawled Multilingual Datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75, Melbourne, Australia. Association for Computational Linguistics.

Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, Joe Davison, Mario Šaško, Gunjan Chhablani, Bhavitvya Malik, Simon Brandeis, Teven Le Scao, Victor Sanh, Canwen Xu, Nicolas Patry, Angelina McMillan-Major, Philipp Schmid, Sylvain Gugger, Clément Delangue, Théo Matussière, Lysandre Debut, Stas Bekman, Pierric Cistac, Thibault Goehringer, Victor Mustar, François Lagunas, Alexander Rush, and Thomas Wolf. 2021. Datasets: A community library for natural language processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 175–184, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *Proceedings of the Seventh International Conference on Learning Representations (ICLR'19)*.

Alessio Miaschi, Gabriele Sarti, Dominique Brunato, Felice Dell'Orletta, and Giulia Venturi. 2021. Italian transformers under the linguistic lens. In *Proceedings of the Seventh Italian Conference on Computational Linguistics (CLiC-it 2020)*, Online. CEUR.org.

El Moatez Billah Nagoudi, AbdelRahim Elmadany, and Muhammad Abdul-Mageed. 2021. AraT5: Text-to-text transformers for arabic language understanding and generation. *ArXiv Computation and Language*, arXiv:2109.12068(v3).

Debora Nozza, Federico Bianchi, and Dirk Hovy. 2020. What the [MASK]? making sense of language-specific BERT models. *ArXiv Computation and Language*, arXiv:2003.02912(v1).

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Sudha Rao and Joel Tetreault. 2018. Dear sir or madam, may I introduce the GYAFC dataset: Corpus, benchmarks and metrics for formality style transfer. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 129–140, New Orleans, Louisiana. Association for Computational Linguistics.

Arij Riabi, Thomas Scialom, Rachel Keraron, Benoît Sagot, Djamé Seddah, and Jacopo Staiano. 2021. Synthetic data augmentation for zero-shot cross-lingual question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7016–7030, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Phillip Rust, Jonas Pfeiffer, Ivan Vulić, Sebastian Ruder, and Iryna Gurevych. 2021. How good is your tokenizer? on the monolingual performance of multilingual language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3118–3135, Online. Association for Computational Linguistics.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Stella Biderman, Leo Gao, Tali Bers, Thomas Wolf, and Alexander M. Rush. 2021. Multitask prompted training enables zero-shot task generalization. *ArXiv Computation and Language*, arXiv:2110.08207(v2).

Noam Shazeer. 2020. GLU variants improve transformer. *ArXiv Machine Learning*, arXiv:2002.05202(v1).

Noam Shazeer and Mitchell Stern. 2018. Adafactor: Adaptive learning rates with sublinear memory cost. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 4596–4604. PMLR.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.

Yi Tay, Mostafa Dehghani, Jinfeng Rao, William Fedus, Samira Abnar, Hyung Won Chung, Sharan Narang, Dani Yogatama, Ashish Vaswani, and Donald Metzler. 2022. Scale efficiently: Insights from pre-training and fine-tuning transformers. In *Proceedings of the Tenth International Conference on Learning Representations (ICLR'22)*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. 2021a. ByT5: Towards a token-free future with pre-trained byte-to-byte models. *ArXiv Computation and Language*, arXiv:2105.13626.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021b. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *Proceedings of the Eigth International Conference on Learning Representations (ICLR'20)*.

A Generation Examples using IT5 Base

In the following, we use [...] to hide portions of long source texts that we judge less salient to make the examples more readable.

A.1 Wikipedia Summarization (WITS)

Source text:
Volume Entertainment ha prodotto e pubblicato l'audiolibro "Fantozzi", dal primo libro della saga di Paolo Villaggio del 1971, aggiornato e registrato dallo stesso Villaggio nel novembre 2015. [...] Nel 2017 Volume Entertainment produce il docu-film "La voce di Fantozzi", riconosciuta opera di interesse culturale dal MIBACT e presentato in concorso in anteprima alla 74 Mostra internazionale d'arte cinematografica di Venezia nello stesso anno. Volume Audiobooks è la divisione dedicata alla produzione e distribuzione di audiolibri e podcast. [...] Volume Media è specializzata nella produzione di documentari e format TV. [...]

Generated text:
libro entertainment è una casa editrice italiana fondata nel 2015 da paolo villaggio e l'editore e distributore di audiolibri e podcast.

Source text:
Intorno alle 7:15 del mattino del 10 aprile 2002, un militante palestinese si fece esplodere sull'autobus Egged numero 960 che era in viaggio da Haifa a Gerusalemme. [...] Otto persone furono uccise nell'attacco e altre 19 rimasero ferite, tre delle quali gravemente. La maggior parte dei feriti erano soldati e poliziotti israeliani. La forza dell'esplosione danneggiò anche molti veicoli nelle vicinanze. [...]

Generated text:
l'attentato di haifa è stato un attacco terroristico israeliano avvenuto il 10 aprile 2002 ad haifa, in israele.

A.2 News Summarization (NewsSum-IT)

Source text:
Mercoledì 18 settembre il cofondatore e CEO di Google, Larry Page, ha annunciato la creazione di Calico, una nuova società che si occuperà dello sviluppo di soluzioni per la salute e per il benessere. Calico si occuperà principalmente di affrontare i problemi legati all'invecchiamento e ad alcune malattie, come il cancro. [...] Calico sarà principalmente dedicata alla ricerca di cure e terapie contro l'invecchiamento e le malattie degenerative. I piani, come ammette lo stesso Page, non sono ancora del tutto chiari perché il progetto è appena partito e ha obiettivi i cui risultati saranno visibili solo nel lungo periodo.

Generated text:
google ha creato calico, una nuova società che si occuperà di curare il cancro.

Source text:
Clubhouse, un nuovo social network in cui invece di scrivere brevi messaggi o condividere immagini si parla in diretta, sta attirando velocemente le attenzioni di giornali e appassionati di internet anche in Italia, dopo che nelle scorse settimane lo aveva fatto negli Stati Uniti. [...] Esiste dallo scorso marzo, ed è una specie di forum, ma orale invece che scritto. Potrebbe assomigliare a Telegram, se Telegram fosse fatto di soli messaggi vocali, con la differenza che in Clubhouse non c'è nulla di registrato: si parla live, chiedendo di intervenire con una simbolica alzata di mano. [...]

Generated text:
il nuovo social network che parla in diretta. è un forum orale, ma orale invece che scritto, e sta attirando le attenzioni di giornali e appassionati di internet.

A.3 Question Answering (SQuAD-IT QA)

Source text:
La crisi petrolifera del 1973 iniziò nell' ottobre 1973 quando i membri dell' Organizzazione dei Paesi esportatori di petrolio arabo (OAPEC, composta dai membri arabi dell' OPEC più Egitto e Siria) proclamarono un embargo petrolifero. [...] Più tardi fu chiamato il "primo shock petrolifero", seguito dalla crisi petrolifera del 1979, definita il "secondo shock petrolifero". **Domanda: Chi ha proclamato l' embargo petrolifero?**

Generated text:
organizzazione dei paesi esportatori di petrolio arabo

Source text:
Negli Stati Uniti, gli studiosi sostengono che esisteva già un accordo negoziato basato sull'

uguaglianza tra le due parti prima del 1973. La possibilità che il Medio Oriente potesse diventare un altro confronto di superpotenza con l' URSS era più preoccupante per gli Stati Uniti che per il petrolio. Inoltre, i gruppi d' interesse e le agenzie governative più preoccupati per l' energia non erano all' altezza della dominanza di Kissinger. [...] **Domanda: I gruppi d' interesse e le agenzie governative che si occupavano di energia non erano all' altezza di chi?**

Generated text: kissinger

A.4 Question Generation (SQuAD-IT QG)

Source text:
In seguito all' evento di estinzione del Cretaceo-Paleogene, l' estinzione dei dinosauri e il clima umido possono aver permesso alla foresta pluviale tropicale di diffondersi in tutto il continente. Dal 66-34 Mya, la foresta pluviale si estendeva fino a sud fino a 45°. [...] Tuttavia, la foresta pluviale è riuscita ancora a prosperare durante questi periodi glaciali, consentendo la sopravvivenza e l' evoluzione di un' ampia varietà di specie. **Risposta: l' estinzione dei dinosauri e il clima umido**

Generated text:
quali fattori hanno permesso alla foresta pluviale tropicale di diffondersi in tutto il continente?

Source text:
Il primo europeo a percorrere il Rio delle Amazzoni fu Francisco de Orellana nel 1542. [...] Le Storie Unnatural Histories della BBC hanno presentato la prova che la foresta pluviale amazzonica, piuttosto che essere un deserto incontaminato, è stato modellato dall' uomo per almeno 11.000 anni attraverso pratiche come il giardinaggio forestale e la terra preta. **Risposta: Francisco de Orellana**

Generated text:
chi è stato il primo europeo a percorrere il rio delle amazzoni?

A.5 Headline Style Transfer (CHANGE-IT)

Source text:
Il dopo Renzi: Padoan in pole, ipotesi Grasso. Le carte in mano a Mattarella

Generated text:
mattarella e il piano b: il premier al quirinale

Source text:
Molise, centrodestra vince e bussa al Colle. Berlusconi: "M5s battuti, il governo spetta a noi"

Generated text:
elezioni molise, il m5s non ce la fa

Source text:
Insulta i capi in una chat privata. Ma il collega lo tradisce e ora è inquisito

Generated text:
trieste, sindacalista ugl insulta due dirigenti su whatsapp

A.6 Headline Generation (HeadGen-IT)

Source text:
C'è chi perde la casa, tra una scossa e l'altra del terremoto, e vive un doppio dramma perché sfollato con un quattrozampe al seguito. Lo denuncia l'Enpa (Ente nazionale protezione animali) spiegando che, in alcune strutture di accoglienza destinate ai terremotati del centro Italia, gli animali domestici non sarebbero graditi. [...] Sono circa quindici i cani e i gatti tratti in salvo nelle ultime ore tra Norcia e le altre località del terremoto; la stima attuale è che 5mila tra cani e gatti abbiano seguito i loro padroni nelle tendopoli o negli alberghi messi a disposizione. [...]

Generated text:
animali al seguito rifiutati dagli albergatori appello dell'enpa: "anche loro sono sfollati"

Source text:
Parla della crisi del Venezuela, Papa Francesco: "Mi fa paura lo spargimento di sangue. E per questo chiedo di essere grandi a coloro che possono aiutare a risolvere il problema. Il problema della violenza mi atterrisce. E se hanno bisogno di aiuto che si mettano dáccordo e lo chiedano". [...] Oltre ai pastori anche i cristiani, i cattolici ipocriti, che vanno tutte le domeniche a messa e poi non pagano la tredicesima, pagano in nero, sfruttano la gente, poi vanno ai Caraibi a fare le vacanze. Ḿa io sono cattolico, vado tutte le domeniche a messa!, Se tu fai questo dai una contro testimonianza. [...]

Generated text:
il papa: "evitare la violenza in venezuela". e sul lavoro: "ipocriti i cattolici che lo sfruttano"

A.7 Formal-to-Informal Style Transfer (XFORMAL IT F $\rightarrow$ I)

Source text:
evita di opprimerlo eccessivamente.

Generated text:
non opprimerlo troppo.

Source text:
esprimile ad alta voce l'amore che provi nei suoi confronti.

Generated text:
dille quanto ti piace.

Source text:
sto facendo in modo di attivarmi per una richiesta ricevuta da lunaurora, dunque partecipo sicuramente alla chat.

Generated text:
io lo faccio per una richiesta lunaurora e partecipo sicuramente alla chat.

A.8 Informal-to-Formal Style Transfer (XFORMAL IT I $\rightarrow$ F)

Source text:
ciao lo so che nn centra nulla ma nn so propio come contattarti!!

Generated text:
ciao, so che non c'entra nulla, ma non so come contattarti.

Source text:
concordo decisamente con verdina b xò nn t devi far vedere da lui!cm è 1 idea bellissima quella di verdina b!

Generated text:
concordo con verdina b, ma non devi farti vedere da lui.

	IT5 Small	IT5 Base	IT5 Large
# of parameters	60M	220M	738M
# of steps	1'050k	1'050k	2'100k
Training time	36 h	101 h	370 h
Batch size	128	128	64
Weight decay	1e-3	1e-3	1e-2
Feedforward size	1024	2048	2816
Hidden size	512	768	1024
# encoder layers	6	12	24
# decoder layers	6	12	24
# attention heads	6	12	16
K-V proj. size		64	
Dropout rate		0	
Non-linearity		Gated GeLU	
LayerNorm ϵ		1e-6	
# rel. att. buckets		32	
Vocabulary size		32'000	

Table 5: Full parametrization for IT5 models. Parameters below the line are shared across all configurations.

Source text:
meglio 1 pò di pancetta e tanta allegria ke 1 triste e insoddisfatta ragazza pelle e ossa!

Generated text:
e' meglio avere un po' di pancetta e tanta allegria che una triste e insoddisfatta ragazza pelle e ossa.

B IT5 Full Model Parametrization

Table 5 shows the full parameter configuration for the three IT5 model sizes. The models correspond to the three canonical sizes for T5 models, with T5 v1.1 improvements.

C Italian Baseline Scores for BERTScore Rescaling

Table 7 contains the baseline scores computed on the first 1M examples of the Cleaned Italian mC4 Corpus using the same model, which we later use for evaluating generation performances. These should be used alongside the same model and the `-rescale_with_baseline` option to obtain BERTScore performances directly comparable to the ones reported in this work.

The hash code used for reproducibility by the BERTScore library is `dbmdz/bert-base-italian-xxl-uncased_L10_no-idf_version=0.3.11(hug_trans=4.16.0)-rescaled`

	WITS	NewsSum	QA	QG	XFORMAL F $\leftrightarrow$ I	CHANGE	HeadGen
Max. source length	1000	512	512	512	128	512	512
Max target length	128	128	64	128	64	64	64
Num. epochs	3	7	7	7	10	10	7

Table 6: Task-specific fine-tuning parameters.

Layer	Precision	Recall	F1
0	0.3164	0.3165	0.3100
1	0.3869	0.3870	0.3843
2	0.3777	0.3778	0.3759
3	0.4955	0.4955	0.4945
4	0.5646	0.5646	0.5637
5	0.5874	0.5874	0.5868
6	0.5712	0.5713	0.5706
7	0.5483	0.5484	0.5478
8	0.4989	0.4989	0.4979
9	0.4401	0.4401	0.4382
10	0.4082	0.4082	0.4061
11	0.3766	0.3766	0.3750
12	0.3400	0.3400	0.3381

Table 7: Baseline scores for using `dbmdz/bert-base-italian-xxl-uncased` with the BERTScore evaluation framework.

D Parametrization for Fine-tuning Experiments

Table 6 contains task-specific parameters that were used for the fine-tuning experiments. For mT5 Small, IT5 Small and IT5 Base models we use a learning rate of 5e-4 and a batch size of 64 examples, while larger models (mT5 Base and IT5 Large) were fine-tuned with a leaning rate of 5e-5 and a batch size of 32. All models are fine-tuned with linear schedule with no warmup using the AdamW optimizer (Loshchilov and Hutter, 2019).

We highlight that the batch sizes used for fine-tuning are significantly smaller from the canonical batch size of 128 adopted by Raffel et al. (2020) due to hardware limitations.