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

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

Abstract

We introduce IT5, the first family of encoder-decoder transformer models pretrained specifically on Italian. We document and perform a thorough cleaning procedure for a large Italian corpus and use it to pretrain four IT5 model sizes. We then introduce the ItaGen benchmark, which includes a broad range of natural language understanding and generation tasks for Italian, and use it to evaluate the performance of IT5 models and multilingual baselines. We find 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 Italian language generation.

Keywords: Italian NLP, Natural Language Generation, Language Modeling

1. Introduction

The text-to-text paradigm introduced by T5 (Raffel et al., 2020) has been widely adopted as a simple yet powerful generic transfer learning approach for most language processing tasks (Sanh et al., 2022; 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., 2022, 2021), 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., 2023). Multilingual language models were shown to excel in cross-lingual and low-resource scenarios. Still, multiple studies have highlighted their suboptimal scale-to-performance ratio when compared to monolingual counterparts for language-specific applications in which data are abundant (Nozza et al., 2020; Rust et al., 2021).

For this reason, monolingual T5 models have recently been pretrained to serve specific language communities, covering languages such as Arabic, Portuguese, Vietnamese and Slovenian (Nagoudi et al., 2022; Carmo et al., 2020; Phan et al., 2022; Ulčar and Robnik-Šikonja, 2023). These models improve over multilingual baselines on language understanding and generation tasks such as news summarization, headline generation and natural language inference in their respective languages.

In this work, we follow a similar approach to pretrain and evaluate four Italian T5 models of different sizes, which we identify as **IT5**. In Section 3, we present the cleaning procedure for the Italian portion of the mC4 corpus (Xue et al., 2021) used in pre-training the IT5 models. Section 4 describes multilingual baselines and downstream tasks used to evaluate fine-tuned IT5 models and presents the obtained results. Finally, our findings and future directions are summarized in Section 5. We make the following contributions:

- We introduce a large-scale cleaned version of the Italian mC4 corpus and use it to pre-train four IT5 models of various dimensions.
- We introduce ItaGen, a benchmark for Italian language understanding and generation tasks.
- We evaluate IT5 on ItaGen, showing improvements over multilingual baselines and previous state-of-the-art work.
- We publicly release all the code, data, and pre-trained/fine-tuned checkpoints for further experimentation by the research community.

To the best of our knowledge, our IT5 models are the first publicly available encoder-decoder models pre-trained exclusively on the Italian language. IT5 constitutes a significant contribution to Italian NLP, as evidenced by its prompt adoption by the research community upon its release (La Quatra and Cagliero, 2023; Papucci et al., 2022; Mousavi et al., 2023 *inter alia*), especially in the context of the latest evaluation campaign of Italian NLP tools (Leonardelli and Casula, 2023; Hromei et al., 2023). This paper serves as the prime reference for IT5, providing all relevant details regarding training data and parameters, a battery of experiments on a collection of tasks, which can be further used as a reference benchmark especially for Italian generation, and a discussion of its limitations.[1]

2. Related Work

2.1. Text-to-text Transfer Transformers

The Text-to-text Transfer Transformer (T5) model (Raffel et al., 2020) adapts the original

[1] Resources: `https://github.com/gsarti/it5`.

LREC-COLING 2024, pages 9422–9433
20-25 May, 2024.

Transformer architecture proposed by Vaswani et al. (2017) by reformulating multiple natural language processing tasks into a unified text-to-text format and using them 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 applied to natural language understanding tasks by using a text-to-text format, making the T5 model highly versatile in most NLP settings.

2.2. Pre-trained Language Models for Italian

The high technical expertise and heavy computational resources required for developing state-of-the-art models recently exacerbated inequalities in access to state-of-the-art systems for non-English languages. Despite the good amount of linguistic resources currently available, the Italian NLP community can currently count on a small set of publicly available pre-trained language models based mostly on the BERT architecture – AlBERTo (Polignano et al., 2019), UmBERTo[2] and GilBERTo[3] *inter alia*, see Miaschi et al. (2022) for a survey – and most notably on a single decoder-only model for text generation, GePpeTto (De Mattei et al., 2020a). Our IT5 models fill the current gap in the availability of sequence-to-sequence models, providing natural choices for monolingual tasks such as summarization, question answering and reformulation.

3. Data and Model Pretraining

The original T5 models were pre-trained on the 750GB web-scraped English C4 corpus (Raffel et al., 2020). C4 authors cleaned the corpus with heuristics to remove 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., 2021), a multilingual version of C4 including 107 languages. While mC4 authors adopted a similar procedure, the language detection threshold is lowered to 70% and other useful heuristics are omitted due to their brittleness across various character systems. As a result, the resulting corpus has an overall lower quality, with recent work finding 16% of sampled mC4 examples having the wrong language tag, and 11% not containing any linguistic information (Kreutzer et al., 2022). For this reason, we perform a thorough cleaning of the Italian portion of mC4 before pre-training IT5.

Task	**Dataset**
Wiki Summarization	WITS (Casola and Lavelli, 2022)
News Summarization	NewsSum-IT (Landro et al., 2022)
Question Answering Question Generation	SQuAD-IT (Croce et al., 2018)
Headline Style Transfer Headline Generation	CHANGE-IT (De Mattei et al., 2020b)
Formality Style Transfer	XFORMAL-IT (Briakou et al., 2021)

Table 1: Summary of datasets composing ItaGen.

3.1. Cleaning the Italian mC4 Corpus

The original Italian mC4 Corpus includes approximately 359GB of raw text data and is one of the largest public Italian corpora. To perform a more thorough cleaning of this data, we use a public implementation[4] reproducing and improving the original C4 data cleaning pipeline. Specifically, we sentence-tokenize documents and remove sentences containing either (i) words from a manually selected subset of the Italian and English List of Dirty Naughty Obscene and Otherwise Bad Words;[5] (ii) less than three words, or a word longer than 1000 characters; (iii) an end symbol not matching standard end-of-sentence punctuation for Italian; or (iv) strings associated to Javascript code, lorem ipsum, English and Italian privacy policy/cookie disclaimers. We finally keep only documents containing more than five sentences, having between 500 and 50k characters, and having Italian as the main language.[6] The resulting Clean Italian mC4 Corpus[7], contains roughly 215GB of raw Italian text, corresponding roughly to 103M documents and 41B words.

3.2. Model and Training Parameters

The first 10M documents sampled from the cleaned corpus are used to train a SentencePiece unigram subword tokenizer (Kudo, 2018) with a vocabulary size of 32k words. The full cleaned corpus is then used to pre-train three models following the canonical small, base and large sizes (Raffel et al., 2020). Moreover, a fourth model adopting the efficient small EL32 architecture by Tay et al. (2022) is also pre-trained and evaluated.

[2] https://github.com/musixmatchresearch/umberto

[3] https://github.com/idb-ita/GilBERTo

[4] https://gitlab.com/yhavinga/c4nlpreproc

[5] https://github.com/LDNOOBW

[6] We use the langdetect toolkit.

[7] https://hf.co/datasets/gsarti/clean_mc4_it

4. Evaluation

4.1. The ItaGen Benchmark

We propose a selection of seven representative tasks, collectively referred to as ItaGen, to evaluate the downstream performances of fine-tuned IT5 and mT5 models. ItaGen aims to provide a comprehensive overview of canonical conditional text generation applications such as summarization, style transfer and question generation, and is constrained by the limited availability of Italian corpora for such tasks. Moreover, we also include a direct comparison of IT5 performances against encoder-based extractive systems for extractive question answering. Table 1 provides an overview of ItaGen tasks and datasets.

Wikipedia Summarization We evaluate encyclopedic summarization on the Wikipedia for Italian Text Summarization (WITS) corpus (Casola and Lavelli, 2022), containing 700k articles extracted from a cleaned dump of the Italian Wikipedia alongside their leading sections used as approximated summaries. We adopt the original evaluation setup using a 10k examples test set.

News Summarization We evaluate news article summarization by concatenating Fanpage.it and IlPost newspapers articles collected by Landro et al. (2022). We refer to this concatenated corpus as NewsSum-IT. We fine-tune our systems on the training set, including roughly 100k 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 extractive question answering using the SQuAD-IT dataset (Croce et al., 2018), containing 50k paragraph-question-answers triplets automatically translated from the English SQuAD dataset (Rajpurkar et al., 2016). We frame the QA task as a text-to-text problem aimed at generating responses given a source text using the format `<CONTEXT> Domanda: <QUESTION>`. We use the original evaluation script and splits.

Question Generation We use SQuAD-IT to evaluate question generation capabilities by reordering the text triplets, making the model predict a plausible question given a source text in the format `<CONTEXT> Risposta: <ANSWER>`, where the answer is the first among the available per-example answers, using the same train-test splits of QA.

Headline Style Transfer We evaluate style transfer abilities in the news domain on the CHANGE-IT shared task (De Mattei et al., 2020b), containing 60k newspaper articles and headlines from the left-leaning Italian newspaper la Repubblica and the right-leaning Il Giornale, respectively. We train and validate our models on author-defined splits using the original cross-source article-to-headline generation. We report average scores for the two style transfer directions (Il Giornale ↔ la Repubblica).

Headline Generation To evaluate news headline generation, we combine the two CHANGE-IT subsets to create a corpus of roughly 120k news articles and headlines pairs, which we refer to with the name HeadGen-IT. Original CHANGE-IT test sets are preserved.

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 115k 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 crowd workers via Amazon Mechanical Turk. We evaluate our models in both style transfer directions (Formal ↔ Informal).

4.2. Evaluation Metrics

We use a combination of common lexical and trainable metrics across all available tasks. We use the language-independent ROUGE metric (Lin, 2004) in its R1, R2 and RL variants to evaluate lexical matches, and BERTScore (Zhang et al. 2020; BS) to evaluate correspondence at the semantic level.[8] For QA, the canonical exact-match (EM) and F1-score (F1) metrics are used. Finally, for the news headline style transfer task, we use trained classifiers provided by the authors[9] to ensure comparable headline-headline (HH) and headline-article (HA) coherence performances.

4.3. Baselines

Besides baselines available from previous studies using the selected datasets, we also adopt the same fine-tuning procedure for fine-tuning two sizes (small and base) of the multilingual T5 model (mT5) (Xue et al., 2021). These multilingual models are used to assess the validity of our pre-training procedure and to observe whether the monolingual setting improves performance.[10]

[8] We use an Italian BERT model to obtain baseline scores to broaden the metric range and remove noise, following authors' recommendations.

[9] `michelecafagna26/CHANGE-IT`

[10] mT5 models are bigger than T5s due to larger vocabularies and embedding matrices, making their usage on consumer accelerators more challenging.

	WITS			
	R1	R2	RL	BS
TextRank (2022)	.302	.076	.197	-
LexRank (2022)	.269	.059	.175	-
SumBasic (2022)	.206	.048	.140	-
mT5 Small	.347	.200	.316	.517
mT5 Base	.348	.200	.315	.520
IT5 Small (ours)	.337	.191	.306	.504
IT5 EL32 (ours)	.346	.196	.314	.513
IT5 Base (ours)	**.369**	**.217**	**.333**	**.530**
IT5 Large (ours)	.335	.191	.301	.508

	CHANGE-IT			
	HH	HA	RL	BS
PointerNet (2020b)	.644	.874	.151	-
BiLSTM$_{+\mathtt{Att}}$ (2020c)	.744	.846	.155	-
mT5 Small	.777	.807	.211	.372
mT5 Base	.795	.799	.236	.398
IT5 Small (ours)	.898	**.882**	.231	.392
IT5 EL32 (ours)	.822	.786	.244	.406
IT5 Base (ours)	**.904**	.868	**.247**	**.411**
IT5 Large (ours)	.895	.861	.237	.390

	Size	SQuAD-IT QA	
	#	F1	EM
DrQA-IT (2018)	-	.659	.561
mBERT (2019)	110M	.760	.650
BERT-IT [11] (2019)	110M	.753	.638
XLM-R Large$_{+\mathtt{st}}$ (2021)	560M	**.804**	.676
mT5 Small	300M	.660	.560
mT5 Base	580M	.757	.663
IT5 Small (ours)	60M	.716	.619
IT5 EL32 (ours)	143M	.747	.645
IT5 Base (ours)	220M	.761	.663
IT5 Large (ours)	738M	.780	**.691**

	NewsSum-IT				SQuAD-IT QG				HeadGen-IT				XFORMAL-IT F→I				XFORMAL-IT I→F			
	R1	R2	RL	BS	R1	R2	RL	BS	R1	R2	RL	BS	R1	R2	RL	BS	R1	R2	RL	BS
mBART Large (2022)	.323	.150	.248	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
mT5 Small	.340	.161	.262	.375	.306	.143	.286	.463	.277	.094	.244	.408	.651	**.450**	.631	.666	.638	.446	.620	.684
mT5 Base	.330	.155	.258	.393	.346	.174	.324	.495	.302	.109	.265	.427	**.653**	.449	**.632**	.667	.661	.471	.642	.712
IT5 Small (ours)	.354	.172	.278	.386	.367	.189	.344	.505	.287	.100	.253	.414	.650	**.450**	.631	.663	.646	.451	.628	.702
IT5 EL32 (ours)	.339	.160	.263	**.410**	.382	.201	.357	.517	.299	.108	.264	.427	.459	.244	.435	**.739**	.430	.221	.408	.630
IT5 Base (ours)	.251	.101	.195	.315	.382	.199	.354	.516	**.310**	.112	**.270**	**.433**	.652	.446	.632	.665	.583	.403	.561	.641
IT5 Large (ours)	**.377**	**.194**	**.291**	-	**.383**	**.204**	**.360**	**.522**	.308	**.113**	**.270**	.430	.611	.409	.586	.613	**.663**	**.477**	**.645**	**.714**

Table 2: IT5, mT5 and baseline models performances on ItaGen datasets. Best scores are highlighted.

4.4. Results and Discussion

Table 2 present the results of our fine-tuning experiments. Given the broad scope of our analysis, we limit ourselves to comment on salient trends we observe across tasks and model categories.

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, particularly for question answering and generation and for headline-headline coherence on the news headline style transfer task. For QA, the IT5 Large model outperforms most extractive systems, including the XLM-R Large by Riabi et al. (2021), despite its adhoc synthetic data augmentation procedure.

Multilingual models can still be helpful in specific applications and when using translated data. We observe that multilingual language models perform best 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, after a preliminary 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 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 common scaling trends for Transformers (Brown et al., 2020), we do not observe a systematic increase in downstream performances for IT5 models when increasing their size, despite lower loss scores and higher accuracies achieved by larger models during pre-training. 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 our results might be related to a bottleneck in the maximal batch size for large models, set to 128 examples instead of the 2048 reported by Raffel et al. (2020) due to lack of resources. We observe that the EL32 architecture can frequently outperform larger model variants, suggesting efficient model design as a promising direction for monolingual model development.

5. Conclusion

This paper introduced IT5, the first family of large-scale encoder-decoder models pre-trained in Italian. We presented a detailed overview of the overall training and evaluation procedure, including comparisons with multilingual counterparts on a broad set of Italian language generation tasks. 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 deem a further in-

[11] bert-base-italian-uncased-squad-it

vestigation of time and quality trade-offs between pre-trained monolingual models and a language-specific continued pre-training of multilingual models as a future step to further narrow the gap in modeling performances for less-resourced languages.

6. Acknowledgements

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.

7. Ethics 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, we encourage further studies to assess the magnitude and prevalence of such biases. Model usage should ideally be restricted to research-oriented and non-user-facing endeavors.

Due to our limited computational resources, we could not conduct an exhaustive hyperparameter search for pre-training and fine-tuning IT5 models. For this reason, reported scores should not be treated as the best achievable results, and further improvements can undoubtedly be achieved with additional benchmarking effort.

Finally, despite using standard metrics capturing lexical and semantic similarity to evaluate our models, we do not explicitly evaluate the factual consistency of generated outputs. For this reason, the real-world effectiveness of our models should be further assessed in future studies, especially for tasks prone to hallucination, such as abstractive summarization.

8. Bibliographical 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)*, Online. OpenReview.net.

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'21)*, 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'19)*, 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'20)*, 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. 2023. DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *Proceedings of the Eleventh International Conference on Learning Representations (ICLR'23)*, Kigali, Rwanda. OpenReview.net.

Claudiu D. Hromei, Danilo Croce, Valerio Basile, and Roberto Basili. 2023. ExtremITA at EVALITA 2023: Multi-task sustainable scaling to large language models at its extreme. In *Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023)*, volume 3473 of *CEUR Workshop Proceedings*, Parma, Italy. CEUR-WS.org.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12).

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.

Moreno La Quatra and Luca Cagliero. 2023. Bart-it: An efficient sequence-to-sequence model for italian text summarization. *Future Internet*, 15(1).

Nicola Landro, Ignazio Gallo, Riccardo La Grassa, and Edoardo Federici. 2022. Two new datasets for italian-language abstractive text summarization. *Information*, 13(5).

Elisa Leonardelli and Camilla Casula. 2023. DH-FBK at HODI: multi-task learning with classifier ensemble agreement, oversampling and synthetic data. In *Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023)*, volume 3473 of *CEUR Workshop Proceedings*, Parma, Italy. CEUR-WS.org.

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)*, New Orleans, LA, USA. OpenReview.net.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

Alessio Miaschi, Gabriele Sarti, Dominique Brunato, Felice Dell'Orletta, and Giulia Venturi. 2022. Probing linguistic knowledge in italian neural language models across language varieties. *Italian Journal of Computational Linguistics (IJ-CoL)*, 8(1):25–44.

Seyed Mahed Mousavi, Simone Caldarella, and Giuseppe Riccardi. 2023. Response generation in longitudinal dialogues: Which knowledge representation helps? In *Proceedings of the 5th Workshop on NLP for Conversational AI (NLP4ConvAI 2023)*, pages 1–11, Toronto, Canada. Association for Computational Linguistics.

El Moatez Billah Nagoudi, AbdelRahim Elmadany, and Muhammad Abdul-Mageed. 2022. AraT5: Text-to-text transformers for Arabic language generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 628–647, Dublin, Ireland. Association for Computational Linguistics.

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).

Michele Papucci, Chiara De Nigris, Alessio Miaschi, and Felice Dell'Orletta. 2022. Evaluating text-to-text framework for topic and style classification of italian texts. In *Proceedings of the Sixth Workshop on Natural Language for Artificial Intelligence (NL4AI 2022)*, volume 3287 of *CEUR Workshop Proceedings*, pages 56–70, Udine, Italy. CEUR-WS.org.

Long Phan, Hieu Tran, Hieu Nguyen, and Trieu H. Trinh. 2022. ViT5: Pretrained text-to-text transformer for Vietnamese language generation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop*, pages 136–142, Hybrid: Seattle, Washington + Online. Association for Computational Linguistics.

Marco Polignano, Pierpaolo Basile, Marco de Gemmis, Giovanni Semeraro, and Valerio Basile. 2019. Alberto: Italian BERT language understanding model for NLP challenging tasks based on tweets. In *Proceedings of the Sixth Italian Conference on Computational Linguistics (CLiC-it '19)*, volume 2481 of *CEUR Workshop Proceedings*, Bari, Italy. CEUR-WS.org.

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, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal V. 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 Févry, Jason Alan Fries, Ryan Teehan, Teven Le Scao, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M. Rush. 2022. Multitask prompted training enables zero-shot task generalization. In *Proceedings of the Tenth International Conference on Learning Representations (ICLR'22)*, Online. OpenReview.net.

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 (ICML'18)*, 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)*, Online. OpenReview.net.

Matej Ulčar and Marko Robnik-Šikonja. 2023. Sequence-to-sequence pretraining for a less-resourced slovenian language. *Frontiers in Artificial Intelligence*, 6.

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.

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. 2022. ByT5: Towards a token-free future with pre-trained byte-to-byte models. *Transactions of the Association for Computational Linguistics*, 10:291–306.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. 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 Eighth International Conference on Learning Representations (ICLR'20)*, Addis Abeba, Ethiopia. OpenReview.net.

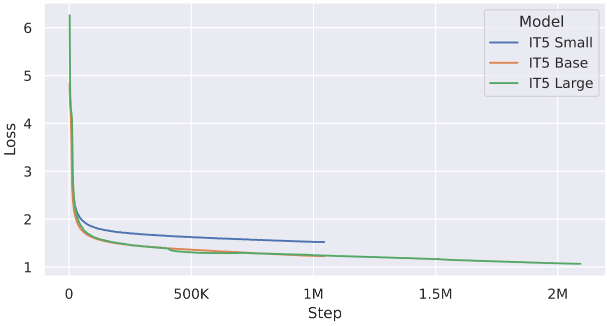

Figure 1: Loss curves for the masked span prediction task used to pre-train the IT5 models.

A. Model Parametrization and Additional Pretraining Details

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[12] also used by the mT5 model, improving upon the original T5 by using GeGLU non-linearities (Shazeer, 2020), scaling model hidden size alongside feedforward 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 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 standard models (excluding the efficient one). 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.[13]

Table 3 shows the full parameter configuration for all IT5 model sizes. The models correspond to the three canonical sizes for T5 models (Small, Base, Large) with T5 v1.1 improvements, and the efficient small version with 32 encoder layers introduced by Tay et al. (2022) (EL32).

B. Italian Baseline Scores for BERTScore Rescaling

Table 4 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.

	Small	Base	Large
# of parameters	60M	220M	738M
# of steps	1.05M	1.05M	2.1M
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 3: Full parametrization for IT5 models. Parameters below the line are shared across all configurations.

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`

C. Parametrization for Fine-tuning Experiments

Table 5 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.

D. Generation Examples using IT5 Base

Tables 6 and 7 present some generation examples for the IT5 Base model across all the evaluated tasks. We use [...] to omit portions of long sources that we judge to be less salient to improve the readability of the examples. Outputs are lowercase

[12] `text-to-text-transfer-transformer/t511`

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

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

Dataset	SL	TL	# Epochs
WITS	100	128	3
NewsSum-IT	512	128	7
SQuAD-IT QA	512	64	7
SQuAD-IT QG	512	128	7
XFORMAL F ↔ I	128	64	10
CHANGE-IT	512	64	10
HeadGen-IT	512	64	7

Table 5: Task-specific fine-tuning parameters. SL = Max. source length. TL = Max. target length.

because the IT5 Base tokenizer is uncased, while using the EL32 model would produce results with normal casing. Examples shown were randomly sampled among model generations for the respective test sets.

While the quality is generally high, we observe that summarization results, especially for the WITS dataset, tend to contain hallucinated information obtained by combining unrelated portions of the source. For example, "Libro Entertainment" in the first example appears to be a translated version of the actual name of the publishing house, and Paolo Villaggio published an audiobook with the company rather than owning it, as it is stated in the generated summary. This is a well-known problem of abstractive summarization systems (Maynez et al., 2020; Ji et al., 2023), which hasn't been studied extensively for languages other than English.

Task	Field	Examples
Wikipedia Summarization (WITS)	Wikipedia Article (Source)	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. [...]
	Summary (Generated)	libro entertainment è una casa editrice italiana fondata nel 2015 da paolo villaggio e l’editore e distributore di audiolibri e podcast.
	Wikipedia Article (Source)	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. [...]
	Summary (Generated)	l’attentato di haifa è stato un attacco terroristico israeliano avvenuto il 10 aprile 2002 ad haifa, in israele.
News Summarization (NewsSum-IT)	News Article (Source)	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.
	Summary (Generated)	google ha creato calico, una nuova società che si occuperà di curare il cancro.
	News Article (Source)	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. [...]
	Summary (Generated)	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.
Question Answering (SQuAD-IT QA)	Context + Question (Source)	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?*
	Answer (Generated)	organizzazione dei paesi esportatori di petrolio arabo
	Context + Question (Source)	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?*
	Answer (Generated)	kissinger
Question Generation (SQuAD-IT QG)	Context + Answer (Source)	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*
	Question (Generated)	quali fattori hanno permesso alla foresta pluviale tropicale di diffondersi in tutto il continente?
	Context + Answer (Source)	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*
	Question (Generated)	chi è stato il primo europeo a percorrere il rio delle amazzoni?

Table 6: Examples for the summarization and question answering/generation tasks from the respective test sets using the IT5 Base model.

Task	Field	Examples
Headline Style Transfer (CHANGE-IT)	La Repubblica Article (Source)	ROMA. E adesso tocca all'arbitro. Da settimane ogni passo di Sergio Mattarella è orientato alla stabilità, ma lo schiaffo referendario sembra rendere impossibile un governo saldo con Matteo Renzi a Palazzo Chigi. Intendiamoci: per il Colle è sempre l'attuale capo dell'esecutivo a dover decidere se restare in sella. Non sarebbe certo il Capo dello Stato, insomma, a non favorire un "Renzi bis". Ma sono la politica, e lo stesso leader, ad essere già passati al piano B. [...]
	Il Giornale Headline (Gen.)	mattarella e il piano b: il premier al quirinale
	La Repubblica Article (Source)	CAMPOBASSO - La prima regione "stellata" non c'è. Davide non ce l'ha fatta contro Golia, una sola lista non riesce a travolgerne 9 dello schieramento avversario. Così il M5s non replica fino in fondo l'onda del 4 marzo e il Molise va al centrodestra, e al suo candidato governatore, il sessantenne commercialista di Fi DonatoToma. [...]
	Il Giornale Headline (Gen.)	elezioni molise, il m5s non ce la fa
Headline Generation (HeadGen-IT)	News Article (Source)	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. [...]
	Headline (Generated)	animali al seguito rifiutati dagli albergatori appello dell'enpa: "anche loro sono sfollati"
	News Article (Source)	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. 'Ma io sono cattolico, vado tutte le domeniche a messa!'. Se tu fai questo dai una contro testimonianza. [...]
	Headline (Generated)	il papa: "evitare la violenza in venezuela". e sul lavoro: "ipocriti i cattolici che lo sfruttano"
Formal-to-Informal Style Transfer (XFORMAL F → I)	Formal (Source)	evita di opprimerlo eccessivamente
	Informal (Generated)	non opprimerlo troppo
	Formal (Source)	esprimile ad alta voce l'amore che provi nei suoi confronti
	Informal (Generated)	dille quanto ti piace
	Formal (Source)	sto facendo in modo di attivarmi per una richiesta ricevuta da lunaurora, dunque partecipo sicuramente alla chat.
	Informal (Generated)	io lo faccio per una richiesta lunaurora e partecipo sicuramente alla chat.
Informal-to-Formal Style Transfer (XFORMAL I → F)	Informal (Source)	ciao lo so che nn centra nulla ma nn so propio come contattarti!!
	Formal (Generated)	ciao, so che non c'entra nulla, ma non so come contattarti.
	Informal (Source)	concordo decisamente con verdina b xò nn t devi far vedere da lui!cm è 1 idea bellissima quella di verdina b!
	Formal (Generated)	concordo con verdina b, ma non devi farti vedere da lui.
	Informal (Source)	meglio 1 pò di pancetta e tanta allegria ke 1 triste e insoddisfatta ragazza pelle e ossa!
	Formal (Generated)	e' meglio avere un po' di pancetta e tanta allegria che una triste e insoddisfatta ragazza pelle e ossa.

Table 7: Examples of headline style transfer, headline generation and formality style transfer tasks from the respective test sets using the IT5 Base model.