WITS: Wikipedia for Italian Text Summarization

Silvia Casola[1,2], **Alberto Lavelli**[2]
1. Università degli studi di Padova
2. Fondazione Bruno Kessler
scasola@fbk.eu, lavelli@fbk.eu

Abstract

Abstractive text summarization has recently improved its performance due to the use of sequence to sequence models. However, while these models are extremely data-hungry, datasets in languages other than English are few. In this work, we introduce WITS (Wikipedia for Italian Text Summarization), a large-scale dataset built exploiting Wikipedia articles' structure. WITS contains almost 700,000 Wikipedia articles, together with their human-written summaries. Compared to existing data for text summarization in Italian, WITS is more than an order of magnitude larger and more challenging given its lengthy sources. We explore WITS characteristics and present some baselines for future work.

1 Introduction

Automatic text summarization aims at condensing one or more source documents in a shorter output, which contains their most salient information. The underlying task can be framed in two different manners: extractive summarizers select the most relevant segments from the input and produce a summary which is a concatenation of such segments; as a result, the output is a subset of the original text, which the summary follows verbatim. On the other hand, abstractive summarizers aim to encode the whole source into an internal representation from which they generate the summary; thus, they produce a new piece of text that condenses the source without necessarily using its vocabulary and expressions.

Recently, abstractive summarization has attracted a growing interest in the Natural Language Processing (NLP) community. Sequence to sequence models have been increasingly used for the task, with pre-trained encoder-decoder transformers becoming the de facto state of the art for abstractive text summarization. Normally pre-trained in an unsupervised manner, these models are then fine-tuned in a supervised way on the downstream dataset; during fine-tuning, the model learns to generate the summary from the source document.


Wikipedia
enciclopedia multilingue collaborativa, online e gratuita

Wikipedia (pronuncia: vedi sotto) è un'enciclopedia online a contenuto libero, collaborativa, multilingue e gratuita, nata nel 2001, sostenuta e ospitata dalla Wikimedia Foundation, un'organizzazione non a scopo di lucro statunitense.

Lanciata da Jimmy Wales e Larry Sanger il 15 gennaio 2001, inizialmente nell'edizione in lingua inglese, nei mesi successivi ha aggiunto edizioni in numerose altre lingue. Sanger ne suggerì il nome,[1] una parola macedonia nata dall'unione della radice *wiki* al suffisso *pedia* (da *enciclopedia*).

Etimologicamente, Wikipedia significa "cultura veloce", dal termine hawaiano *wiki* (veloce), con l'aggiunta del suffisso *-pedia* (dal greco antico *παιδεία*, *paideia*, formazione). Con più di 55 milioni di voci in oltre 300 lingue,[2] è l'enciclopedia più grande mai scritta,[3][4] è tra i dieci siti web più visitati al mondo[5] e costituisce la maggiore e più consultata opera di riferimento generalista su Internet.[6][7][8]

^ Storia

Figure 1: The lead section (from the Wikipedia' own page), which we consider as the article summary. We use the remaining of the article as the source.

While various datasets for abstractive summarization exist for English, resources in other languages are limited. This paper introduces WITS (Wikipedia for Italian Text Summarization), a large-scale dataset for abstractive summarization in Italian, built exploiting Wikipedia. Taking advantage of the structure of Wikipedia pages, which contain a lead section (Figure 1) – giving an overview of the article's topic –, followed by the full-length article – describing the topic in details –, we create a large and challenging dataset for abstractive summarization in Italian, which we will

make publicly available.

WITS is particularly challenging, given its large source length and its high abstractiveness. In this paper, we describe the dataset, its statistics and characteristics, and report some preliminary experiments that might be used as baselines for future work.

This paper is organized as follows: in Section 2, we describe the state of the art in text summarization, focusing on resources for Italian. We later preset the dataset and its related task (Section 3.1); we describe the data collection and preprocessing process in Sections 3.2 and 3.3. In Section 4, we show our results when summarising the dataset using some existing extractive baseline models. Finally, we draw our conclusions in Section 5.

2 State of the Art

Automatic text summarization has recently attracted increasing attention from the NLP community. However, the majority of the research work still focuses on English.

As a matter of example, out of all the papers published in the Association for Computational Linguistics (ACL) conference in 2021, 46 explicitly refer to summarization in their title; 38 of these dealt with English only, while 7 presented experiments with one or more other languages (including 2 on source code summarization). For reference, only one paper (Mastronardo and Tamburini, 2019) on text summarization (in English) was published at the Italian Conference on Computational Linguistics (CLiC-it) since its first edition, and none experimented with Italian.

In this section, we present the state of the art in abstractive text summarization. We first present the available datasets for the task; then, we discuss some relevant learning models. We focus on the significant gap between English and Italian, for which very few resources exist.

2.1 Datasets for Automatic Text Summarization

A typical dataset for text summarization is composed of some source documents (which needs to be summarized) and their corresponding summaries, used as the gold standard. A minority of datasets (e.g., the DUC 2004 dataset[1]) provide multiple gold standards; however, such datasets tend to be small and are mostly used for evaluation.

In general, summaries exploit a human-written abstract. For example, the CNN/Daily Mail Corpus (Nallapati et al., 2016)[2] leverages a bullet-point summary on the newspapers' websites. A similar rationale is used in datasets constructed from scientific papers (Cohan et al., 2018)[3] or patents (Sharma et al., 2019)[4]. In contrast, Rush et al. (2015)[5] frames the task of news summarization as headline generation.

To the best of our knowledge, WikiLingua (Ladhak et al., 2020)[6] is the only summarization dataset that contains data in Italian. WikiLingua is a cross-lingual dataset for abstractive text summarization built on top of WikiHow. WikiHow contains tutorials on how to perform specific tasks in the form of step-by-step instructions. The dataset constructs a summary by concatenating the first sentence for each step and using the remaining text as the source. WikiLingua contains data in 18 languages, including Italian (50,943 source-summary pairs). Both summaries and sources are relatively short (on average, 44 and 418 tokens, respectively, for the Italian split).

2.2 Models for Abstractive Text Summarization

Abstractive text summarization is one of the most challenging tasks in NLP: it requires very long input understanding (encoding), salient passages finding and constrained text generation. Technically, models for abstractive text summarization are generally sequence-to-sequence: they encode the input and then generate the output through a neural network. While some previous work used Recurrent Neural Networks (Chung et al., 2014), with the possible addition of an encoder-decoder attention mechanism (Chopra et al., 2016), transformer models (Vaswani et al., 2017) have later become pervasive, following a similar trend in many other NLP areas. Using self-attention, these models have proved to be superior to Recurrent

[1] https://duc.nist.gov/duc2004/

[2] https://huggingface.co/datasets/cnn_dailymail

[3] https://huggingface.co/datasets/arxiv_dataset

[4] https://huggingface.co/datasets/big_patent

[5] https://huggingface.co/datasets/gigaword

[6] https://huggingface.co/datasets/wiki_lingua

Neural Networks, as they are able to better deal with long dependencies, a critical task in text summarization.

Following another recent trend in NLP, many summarization models use a transfer-learning approach: after a pre-training phase, in which they are training in an unsupervised way on a huge amount of text, they are fine-tuned for the specific downstream task on a relatively limited amount of supervised data. Summarization models either exploit encoders and decoders previously trained for other tasks or are pre-trained from scratch on a specific objective tailored for summarization. Rothe et al. (2020), for example, leveraged previously existing pre-trained models (BERT in Devlin et al. (2019); ROBERTA in Liu et al. (2019); and GPT-2[7] in Radford et al. (2019)) as encoders or decoders of the sequence-to-sequence summarizer and showed high performance improvement with respect to random initialization. More recently, summarization models (Song et al., 2019; Lewis et al., 2020) have been pre-trained with an objective specific to Natural Language Generation tasks. For example, authors of Pegasus (Zhang et al., 2020) used two objectives: Masked Language Model (Devlin et al., 2019) has been widely used in previous work, and consists in masking a percentage of tokens in text, later predicted using context; Gap Sentences Generation is instead a new pre-training objective, in which a percentage of the original sentences are masked, and the model needs to generate them in accordance to the contest.

Following a shared practice, most summarization models have first been trained and evaluated for English only. In some cases, a subsequent multilingual version of the model was also created (Xue et al., 2021). To the best of our knowledge, few sequence-to-sequence models in Italian exist to date[8], and while they might be fine-tuned for summarization, no full-scale evaluation has been performed yet.

[7]GPT-2 ha also been adapted for Italian. See: De Mattei, L., Cafagna, M., Dell'Orletta, F., Nissim, M., & Guerini, M. 2020. GePpeTto Carves Italian into a Language Model. In CLiC-it 2020

[8]See, for example, IT5-base (`https://huggingface.co/gsarti/it5-base`)

3 WITS

3.1 Task and Rationale

Given a Wikipedia article, we extract the lead section (which we sometimes refer to as "Summary" in the remaining of the paper) and propose the following task:

> Given all article sections, summarize its content to produce its lead section.

The task is rather natural given pages structure. According to the Wikipedia Manual of Style[9], the lead section is, in fact, a high-quality summary of the body of the article. The lead "serves as an introduction to the article and a summary of its most important contents" and "gives the basics in a nutshell and cultivates interest in reading on—though not by teasing the reader or hinting at what follows". Moreover, it should "stand on its own as a concise overview of the article's topic".

As for the content, according to Wikipedia, the lead must define the topic, explaining its importance and the relevant context; then, it must summarize the most prominent points of the article, emphasizing the most important material.

Moreover, the lead should only cover information that is contained in the article: "significant information should not appear in the lead if it is not covered in the remainder of the article". This is particularly relevant for abstractive summarization, as models are more prone to produce summaries that are not factual to the source (often called hallucinations) when they are trained to generate summaries containing information not in the source (Nan et al., 2021). The problem of factuality in abstractive summarization is currently an active area of research, as previous work has shown that up to 30% of generated summaries contain non-factual information (Cao et al., 2018).

Linguistically, the lead "should be written in a clear, accessible style with a neutral point of view". It is worth noting that, in contrast to WikiLingua, where the summary is constructed as a concatenation of sentences from different parts of the articles, the summary in WITS is a stand-alone piece of text, with a coherent discourse structure.

3.2 Data Collection

This section describes the process of data collection and preprocessing.

[9]`https://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style`

	WITS		IT-Wikilingua	
# docs	699,426		50,943	
	Summary	Source	Summary	Source
# sentences (avg)	3.75	33.33	5.01	23.52
# tokens (avg)	70.93	956.66	23.52	418.6
Comp. ratio (avg)	16.14		11.67	

Table 1: Datasets statistics. spacy is used for text and sentence tokenization. The number of tokens and sentences is computed for all documents and then averaged.

We downloaded the latest XML dump of Wikipedia in Italian[10], which contains text only. We used Python and the Gensim library to process the file[11]. The original number of documents was 1,454,884. We applied the following exclusion criteria: we removed pages whose title contains numbers only (as they mostly describe years and contain lists of events and references), lists (titles starting with "Lista d"), pages with summaries with less than 80 characters and articles and pages for which the article is less than 1.5 times longer than the lead.

We then preprocessed the text in the following way: from the summary, we removed the content of parentheses (as they often contain alternative names or names in a different language, which cannot be inferred from the article). For the article, we further excluded the following sections, which are not relevant for our task: Note (Footnotes), Bibliografia (References), Voci correlate (See also), Altri progetti (Other projects), Collegamenti esterni (External links), Galleria di Immagini (Images).

3.3 Dataset Statistics

Table 1 shows some statistics on the dataset and compares WITS with the Italian split of WikiLingua (which we will refer to as IT-WikiLingua).

IT-WikiLingua contains documents from 17,673 WikiHow pages, but some of these pages describe more than one method related to the same topic. For example, the page "How to Reduce the Redness of Sunburn" contains several methods: "Healing and Concealing Sunburns", "Lessening Your Pain and Discomfort", and "Preventing a Sunburn". We consider distinct methods as separate documents, as they can be summarized in isolation. Notice that WITS is more than an order of magnitude larger than IT-Wikilingua.

	WITS		IT-Wikilingua	
	Summary	Source	Summary	Source
PER (avg)	1.13	26.21	0.32	1.05
LOC (avg)	2.03	24.07	0.42	1.39
ORG (avg)	0.60	6.65	0.68	0.37
MISC (avg)	19.68	19.68	0.84	3.07
All (avg)	23.44	76.61	1.65	5.88

Table 2: Named Entities in WITS and IT-WikiLingua.

We computed the number of tokens and the number of sentences through the spaCy it-core-news-lg[12] model. Compared to IT-WikiLingua, documents in WITS contains more tokens both in their summary and in their source (which is more than double in length), making the dataset particularly challenging. Note how the sentences are also more lengthy (thus complex) on average. For example, summaries in WITS contain on average less than 4 sentences, but more than 70 words; in contrast, IT-WikiLingua's summaries consist of more than 5 sentences but contain on average 44 tokens. Not surprisingly, WITS' compression ratio is larger than IT-WikiLingua's and very high in absolute value. Finally, we also notice that the dataset is very rich in named entities. Table 2 reports the Named Entities as extracted with spaCy from WITS and IT-Wikilingua.

4 Baselines

We tested some preliminary non-neural baseline methods on the dataset, reported in Table 3.

All methods reported are unsupervised. Thus, we unsupervisedly obtained the summary from the source and then used the lead as the gold standard for evaluation. We evaluated the summaries using Recall-Oriented Understudy for Gisting Evaluation (ROUGE) (Lin, 2004). ROUGE is an n-gram based, recall-oriented metric for summary quality evaluation. Following previous work (Lloret et al., 2018), we report ROUGE-1 (R-1), ROUGE-2 (R-2), and ROUGE-L (R-L) (recall).

We considered the following baselines:

Lead-3 We extract the first three sentences from the source. Previous work has shown that this baseline is often hard to beat (See et al., 2017), especially in news summarization,

[10] https://dumps.wikimedia.org/itwiki/latest/itwiki-latest-pages-articles.xml.bz2

[11] https://radimrehurek.com/gensim/scripts/segment_wiki.html

[12] https://spacy.io/models/it

which presents an "inverted pyramid" structure and tends to report the most important content at the start.

TextRank (Mihalcea and Tarau, 2004) TextRank is an unsupervised algorithm that extracts the most relevant sentences in the source. The algorithm constructs a graph with sentences as nodes and sentence similarity (in terms of shared vocabulary) as edges. The sentences are then ranked by using the PageRank (Page et al., 1999) algorithm.

LexRank (Erkan and Radev, 2004) LexRank works in a similar way as TextRank. However, instead of computing sentence similarity on normalized shared vocabulary, it uses the cosine similarity of their TF-IDF vectors.

SumBasic (Nenkova and Vanderwende, 2005) SumBasic extracts sentences based on their word probabilities. Specifically, it scores each sentence as the mean of the probability of the words it contains (based on their frequency in the document). Iteratively, the sentence with the best score among the ones containing the most probable word is chosen. The probability of the words in the chosen sentence is then squared to limit redundancy.

IT5-small (Raffel et al., 2020) The Text-to-Text Transfer Transformer (T5) is a pre-trained sequence-to-sequence language model, trained treating both input and output as text strings; the rationale is to use the same models for all NLP tasks, unifying them under the sequence-to-sequence framework. We use a small version of the original model (60 million parameters)[13], pretrained on the Clean Italian mC4 IT[14], the Italian split of the multilingual cleaned version of Common Crawl's Corpus (mC4) (Raffel et al., 2020). We extracted 10,000 summary-source pairs from the dataset for the validation set, and 10,000 for the test set. We trained the model on the rest of the data for 100,000 steps; this account for around 30% of the training data. We trained on two GeForce RTX 2080 GPUs and kept the batch size per GPU to 1. We kept the summary length to 75 tokens, and the source text length to 1000 tokens.

[13] https://huggingface.co/gsarti/it5-small

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

	R-1	R-2	R-L
Lead-3	24.76	5.54	16.54
TextRank	30.20	6.57	19.67
LexRank	26.90	5.91	17.52
SumBasic	20.60	4.80	14.01
IT5-small	21.58	9.69	19.34

Table 3: ROUGE results on WITS.

Results show that the Lead-3 baseline performance is low; this is likely due to the structure of Wikipedia, which contains several thematic sections without a general introduction outside the lead section. Extracting the first sentence(s) from each section would likely produce better results and could be investigated in future work.

In contrast, TextRank is the best non-neural baseline, with a ROUGE-2 score of 6.57; LexRank performs comparably. SumBasic metrics are even lower than those obtained with the Lead-3 baseline, suggesting that a purely frequency-based approach is insufficient given the dataset complexity.

Finally, the neural baseline achieves the best results in terms of ROUGE-2, even if it is relatively small and likely severely under-trained, since only around 30% of the data are used for fine-tuning, due to computational constraints. This suggests that sequence-to-sequence neural models have great potential on the dataset, and should be better investigated in future work. Surprisingly, however, results in terms of ROUGE-1 are instead below most of the other baselines. Future work should investigate this discrepancy.

5 Conclusions

We have presented WITS, the first large-scale dataset for abstractive summarization in Italian. We have exploited Wikipedia's articles' structure to build a challenging, non-technical dataset, with high-quality human-written abstracts. Given the lengthy source documents, the short summaries and the short extractive fragments, the dataset calls for an abstractive approach. In the paper, we have explored some standard non-neural extractive baselines and a neural abstractive baseline. Future work will investigate further neural baselines for

the dataset. Moreover, the dataset can be easily extended applying the procedure described in the paper to more languages, including low-resource ones given Wikipedia structure. We are confident that research in summarization in languages other than English will become more active in the near future and hope that WITS can be a valuable step in this direction.

References

Ziqiang Cao, Furu Wei, Wenjie Li, and Sujian Li. 2018. Faithful to the original: Fact aware neural abstractive summarization. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Sumit Chopra, Michael Auli, and Alexander M. Rush. 2016. Abstractive sentence summarization with attentive recurrent neural networks. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 93–98, San Diego, California, June. Association for Computational Linguistics.

Junyoung Chung, Caglar Gulcehre, Kyunghyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. In *NIPS 2014 Workshop on Deep Learning, December 2014*.

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. A discourse-aware attention model for abstractive summarization of long documents. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 615–621, New Orleans, Louisiana, June. 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, June. Association for Computational Linguistics.

Günes Erkan and Dragomir R. Radev. 2004. LexRank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, 22(1):457–479, December.

Faisal Ladhak, Esin Durmus, Claire Cardie, and Kathleen McKeown. 2020. WikiLingua: A new benchmark dataset for multilingual abstractive summarization. In *Findings of EMNLP, 2020*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online, July. 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, July. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, M. Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *ArXiv*, abs/1907.11692.

Elena Lloret, Laura Plaza, and Ahmet Aker. 2018. The Challenging Task of Summary Evaluation: An Overview. *Language Resources and Evaluation*, 52(1).

C. Mastronardo and F. Tamburini. 2019. Enhancing a text summarization system with ELMo. In *CLiC-it*.

Rada Mihalcea and Paul Tarau. 2004. TextRank: Bringing order into text. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 404–411, Barcelona, Spain, July. Association for Computational Linguistics.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar Gulçehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290, Berlin, Germany, August. Association for Computational Linguistics.

Feng Nan, Ramesh Nallapati, Zhiguo Wang, Cicero Nogueira dos Santos, Henghui Zhu, Dejiao Zhang, Kathleen McKeown, and Bing Xiang. 2021. Entity-level factual consistency of abstractive text summarization. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2727–2733, Online, April. Association for Computational Linguistics.

Ani Nenkova and Lucy Vanderwende. 2005. The impact of frequency on summarization. Technical report, Microsoft Research.

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1999. The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab, November.

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

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.

Sascha Rothe, Shashi Narayan, and Aliaksei Severyn. 2020. Leveraging pre-trained checkpoints for sequence generation tasks. *Transactions of the Association for Computational Linguistics*, 8:264–280.

Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 379–389, Lisbon, Portugal, September. Association for Computational Linguistics.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada, July. Association for Computational Linguistics.

Eva Sharma, Chen Li, and Lu Wang. 2019. BIGPATENT: A large-scale dataset for abstractive and coherent summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2204–2213, Florence, Italy, July. Association for Computational Linguistics.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2019. MASS: masked sequence to sequence pre-training for language generation. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 5926–5936. PMLR.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

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 *NAACL*.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2020. PEGASUS: pre-training with extracted gap-sentences for abstractive summarization. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 11328–11339. PMLR.