# ELiRF-VRAIN at BioNLP Task 1B: Radiology Report Summarization

**Vicent Ahuir**[†], **Encarna Segarra**[†,§], **Lluís-F. Hurtado**[†]
[†]VRAIN: Valencian Research Institute for Artificial Intelligence
Universitat Politècnica de València, Spain
[§]ValgrAI: Valencian Graduate School and Research Network of Artificial Intelligence
{viahes,esegarra,lhurtado}@dsic.upv.es

## Abstract

This paper presents our system at the Radiology Report Summarization Shared Task-1B of the 22nd BioNLP Workshop 2023. Inspired by the work of the BioBART model, we continuously pre-trained a general domain BART model with biomedical data to adapt it to this specific domain. In the pre-training phase, several pre-training tasks are aggregated to inject linguistic knowledge and increase the abstractivity of the generated summaries. We present the results of our models, and also, we have carried out an additional study on the lengths of the generated summaries, which has provided us with interesting information.

## 1 Introduction

Radiology reports are documents that interpret radiological examinations. Usually, a radiology report consists of three sections: (1) a background section that describes general information about the patient and exam, (2) a findings section that presents details of the examination, and (3) an impression section that summarizes the findings against the background. This last section is the most crucial for doctors to make clinical decisions.

Due to the recent success of self-supervised learning, the focus of text summarization research has exhibited a gradual shift from extractive techniques to abstractive techniques. The best-performing abstractive models are BART (Lewis et al., 2020), T5 (Raffel et al., 2020), PEGASUS (Zhang et al., 2020a), and GPT-3 (Brown et al., 2020), being all of them Transformers (Vaswani et al., 2017) pre-trained self-supervisedly as denoising sequence to sequence autoencoders. This kind of approaches allow to pre-train deep architectures to learn vast amounts of general linguistic knowledge from large corpora, that can be transferred to downstream tasks by means of fine-tuning. Almost all of these systems used benchmark datasets compiled from news articles, such as the CNN-DailyMail dataset (CNN-DM) (Hermann et al., 2015) and NEWSROOM (Grusky et al., 2018). However, not so many efforts have been carried out in the biomedical domain.

Language models pre-trained on biomedical corpora may further enhance the performance of current biomedical NLG methods, such BioBERT (Lee et al., 2020) or PubMedBERT (Gu et al., 2021). However, there are very few in-domain generative language models for biomedicine. In (Yuan et al., 2022), authors proposed a biomedical autoregressive generative language model, BioBART, pre-trained on the biomedical corpora. They continuously pre-train BART on PubMed[1] abstracts to achieve biomedical domain adaption only using the text-infilling task. The in-domain BioBART outperforms BART model and sets strong baselines for several NLG tasks.

In the framework of BioNLP workshop, some challenges and shared tasks focusing on summarization were created. MEDIQA 2019 edition focused on question entailment and textual inference and their applications in medical Question Answering (Ben Abacha et al., 2019). MEDIQA 2021 (Ben Abacha et al., 2021) promoted research on summarization for consumer health QA and clinical text. In this edition, the winner system (Dai et al., 2021), based on PEGASUS, employed a domain adaptation strategy by further fine-tuning a small amount of in-domain data to improve generalization and transfer abilities.

## 2 Task Description

The Shared Task-1B of the 22nd BioNLP Workshop 2023 (Delbrouck et al., 2023), focuses on the summarization of radiology reports. The task of the summarization of radiology reports can be defined as follows: given a radiology report with findings and background sections, the goal is to

[1] https://pubmed.ncbi.nlm.nih.gov/



*The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 524–529

generate the impression section. For this shared task, the *Impressions* are only generated from the *Findings* section.

Shared Task 1B was divided into two subtasks. The first one is about generating impressions sections based exclusively on the text report. The second one is summarizing the report based on the text information and the indicators that could be extracted from the attached radiology image. The participants were invited to approach both subtasks but were allowed to participate in one; we chose to participate only in the first subtask.

### 2.1 The Dataset

For the subtask where we participated, a dataset is provided based on MIMIC-III (Johnson et al., 2016) with 79 779 samples from two different radiography modalities and six anatomical parts. This dataset was split into four partitions: `train` (59 320 samples), `validation` (7413), `test` (6526), and `hidden-test` (6531).

| | **Findings** | | **Impressions** | |
|---|---|---|---|---|
| | Sent. | Words | Sent. | Words |
| `train` | 8.80 | 124.90 | 3.91 | 52.26 |
| `validation` | 8.85 | 125.69 | 3.93 | 52.65 |
| `test` | 9.28 | 134.95 | 3.75 | 50.97 |
| `hidden-test` | 10.23 | 155.28 | - | - |

Table 1: Average sentences and words on Findings and Impressions for each partition.

Table 1 details the average number of sentences and words for Findings and Impressions, excluding the Impressions of the `hidden-test` that were not available to participants. On the one hand, we notice that `train` and `validation` have similar lengths for both, Impressions and Findings. On the other hand, the test partitions contain longer Findings, especially the `hidden-test`. Moreover, Impressions are shorter in the `test` partition than in the `train` and `validation` ones; thus, `test` presents a higher compression ratio in its Impressions partition than those of the other two partitions.

### 2.2 System Evaluation

For evaluating the systems, the ViLMedic framework (Delbrouck et al., 2022b) was used. It is a framework that aims to increase results reproducibility for medical tasks, such as medical report summarization. Specifically, the systems were evaluated with the following metrics and scores: ROUGE-L (Lin, 2004), BLEU (Papineni et al., 2002), BERTScore (Zhang et al., 2020b), and RadGraph (Delbrouck et al., 2022a).

## 3 Pre-training Model

Inspired by the work of the BioBART model, we continuously pre-trained a general domain BART model with biomedical data to adapt it to this specific domain. Specifically, our starting point was the architecture and weights of the base version of BART[2], publicly available at the repository of HuggingFace (Wolf et al., 2020).

For the pre-training phase, we followed the methodology used in the News Abstractive Summarization models (NAS) work (Ahuir et al., 2021). In NAS, several pre-training tasks are aggregated to inject linguistic knowledge during the pre-training stage and to increase the abstractivity of the generated summaries. We chose this methodology because we hypothesize that reference impressions are written in a mostly abstractive way. Also, the pre-training method helped to transfer more knowledge to the summarization task, which increased the model's performance in the original work.

For pre-training, we chose data that were as similar as possible to radiology reports in terms of vocabulary and grammar. We selected the following MIMIC datasets available at PhysioNet (Goldberger et al., 2000): note events in MIMIC-III (2 083 180 samples) (Johnson et al., 2016), radiology reports in MIMIC-CXR (128 032) (Johnson et al., 2019), and discharge (331 794) and radiology reports (2 321 355) in MIMIC-IV (Johnson et al., 2023). Additionally, we included Wikipedia articles related to medicine to reinforce the domain vocabulary (97 192).

The base version of the BART model is limited to 1024 input tokens; however, most samples exceeded this size. This fact could lead us to lose valuable training data. To overcome this limitation, we split texts into narrower samples as follows: having the text split by lines and a window of no more than 1000 words, we generated sub-samples

[2] `https://huggingface.co/facebook/bart-base`

that contained at least a new line and filled the window with as many words as possible. With this method, we obtained a dataset of 40 894 042 samples.

Due to infrastructure limitations, we could only train the model for one epoch, which took 12 days with four NVIDIA RTX 3090 graphic cards. The following hyperparameters were used: 4 samples per device, 256 gradient accumulation steps, a learning rate of $5 \times 10^{-5}$ with a linear scheduler, 1% of the epoch for warm-up, and 32 bits of training precision. For the hyperparameters not mentioned, we have used the default values of the 4.23.1 version of `transformers` library of HuggingFace[3].

## 4 Models for the Task

For the downstream task, we obtained three models based on our pre-trained model. The first model (M1) was fine-tuned with the `train` partition of the shared task. The second one (M2) was fine-tuned with the `train` and `validation` partitions. Finally, the third one (M3) was fine-tuned with all the partitions with available references: `train`, `validation`, and `test`.

For the fine-tuning phase, we did a grid search of certain hyperparameters with RayTune (Liaw et al., 2018) through the HuggingFace library. We did 20 trials over the following hyperparameters: `learning_rate` (from $8 \times 10^{-6}$ to $4 \times 10^{-5}$), `num_train_epochs` (10 or 15), and `gradient_accumulation_steps` (2, 4, or 8). Since we wanted to find which models obtained a more balanced performance among the four metrics of the task (ROUGE-L, BLEU, RadGraph-F1, and BertScore-F1), we defined the objective to maximize as the harmonic mean (Ferger, 1931) of these four scores. Finally, the three models were fine-tuned during 15 epochs with an NVIDIA RTX 4090 with the following hyperparameters: 8 samples per device, 4 gradient accumulation steps, and a learning rate of $2.14 \times 10^{-5}$.

For the generation of impressions, we used the *generate* method[4] of HuggingFace. To achieve better performance, we identified certain hyperparameters and performed grid search using the M1 model and the `validation` partition, specifically: `max_length` (60, 70, 80, **90**, or 100), `num_beams` (3, 4, **6**, 8, or 10), and `no_repeat_ngram_size` (3, 4, **5**, 6, 8, or 10). The bolded values maximized the harmonic mean score; thus, we fixed them to generate impressions with ours models.

## 5 Results

| Pt | Md | BL | RL | BS | RG | HM |
|---|---|---|---|---|---|---|
| T | M1 | **17.61** | 30.19 | 53.13 | 31.19 | 28.41 |
| | M2 | 17.41 | 29.57 | 52.24 | 31.40 | 28.10 |
| | G1 | 15.99 | **34.07** | **56.30** | **35.25** | 28.89 |
| | G2 | 17.33 | 33.93 | 55.49 | 34.93 | **29.89** |
| HT | M1 | 16.98 | 30.52 | 54.03 | 31.79 | 28.24 |
| | M3 | 18.06 | 30.19 | 53.94 | 32.58 | 29.04 |
| | G1 | **18.36** | **35.32** | **57.26** | **36.94** | **31.42** |

Table 2: Results on test partitions of our models and those of the groups that achieved the highest score on any of the four measures. For all measures, a higher value means a better performance. M1, M2 and M3 are the three models created with our approach. G1 and G2 are the models that have, at least, a highest value in any measure, without taking into account our models.

Table 2 shows the results of our models (M1, M2, M3) and those of the groups (G1, G2) that reached the highest score, excluding our models, on any of the four scores: BLEU (BL), ROUGE-L (RL), BertScore-F1 (BS), and RadGraph-F1 (RG). The overall performance on the four metrics is reflected by the harmonic mean (HM). The results are divided in two sections: `test` (T) and `hidden-test` (HT). The leaderboard scores were computed by limiting the prediction and the reference to 256 words.

Overall, our models have lower performance than the best systems. In the `test` partition, our best model (M1) averages a 9% lower performance than the other two groups if BLEU is excluded from the count, and 5.3% less when is included, meaning that M1 performs substantially better in BLEU than the other systems. Comparing our models, M2 seems to perform worse than M1, despite being trained with more data. In the case of `hidden-test`, our best model (M3) averages 10.7% lower performance than G1 if BLEU is excluded from the count and 8.43% if not. Comparing the performance of our models, unlike what happened in the `test` partition, M3 performs better than M1. Therefore, the inclusion of the `test` partition in training resulted in more acknowledgment

[3] https://huggingface.co/docs/transformers/v4.23.1/en/main_classes/trainer#transformers.TrainingArguments

[4] https://huggingface.co/docs/transformers/v4.23.1/en/main_classes/text_generation#transformers.generation_utils.GenerationMixin

for the model, probably because of the additional findings types.

## 6 Discussion

When we observe Table 2, two main questions rapidly come to our mind: *Why did our models obtain lower values in all scores but BLEU?*, and *Why the M2 model performed worse than M1, despite being trained with more data?*. Surprisingly, both questions point to a main problem in our models: the length of the generated impressions.

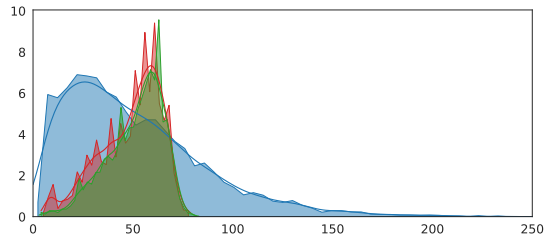


Figure 1: Distribution of samples per number of words. Reference impressions (blue) and generated ones M1 (red), M2 (green). `X-axis`: length of the impressions in words, `Y-axis`: percentage of samples with a certain length.

Figure 1 presents the distribution of the impressions by their length in words. The blue distribution is the reference impressions, the red one is for those generated by M1, and the green one is for M2. It is noticeable that the impressions of our models follow a completely different distribution than the references. However, the three distributions have similar averages in word count: 51 for the references, 49 for the M1, and 51 for the M2. Our models are trying to set a common length for the impressions instead of identifying which ones should be shorter and which ones should be longer. Therefore, BLEU seems to be weaker in this situation than the other metrics. Moreover, M2 generates longer impressions than M1, lowering the precision and, by extension, its general performance. However, there is a chance that M2 excels in some interesting aspects compared to M1.

Table 3 shows the precision and recall obtained by the models M1 and M2 on `test` for ROUGE-L, BERTScore, and RadGraph; also, the harmonic means of these six measures are shown. The SLN group shows the real performance of the models. Contrary, SLY shows the performance when, at most, the first $n$ sentences of the prediction are

| | | precision/recall | | | |
|---|---|---|---|---|---|
| | **M** | **RL** | **BS** | **RG** | **HM** |
| SLN | 1 | **30.59**/36.47 | **52.31**/54.56 | **30.40**/37.43 | **38.17** |
| | 2 | 29.16/**37.55** | 49.66/**55.45** | 30.17/**38.34** | 37.90 |
| SLY | 1 | **35.75**/33.29 | **56.38**/52.77 | **34.05**/34.14 | 39.12 |
| | 2 | 35.73/**33.94** | 56.09/**53.45** | 33.34/**34.69** | **39.26** |

Table 3: Precision and Recall of M1 and M2 models in the `test` partition when there is no sentence limit (SLN) and when the prediction is limited by the number of sentences of the reference (SLY).

taken into account, where $n$ is the number of sentences of the reference. On SLN, we observe that M2 has better recall than M1 but worse precision due to the longer generated impressions, which caused the final lower performance. On SLY, it is noticeable that both models gain more precision than lose recall; thus, our models place more relevant information at the beginning. SLY shows higher harmonic mean values than SLN, which indicates that we could improve the performance of our models by just focusing on making the models increase their focus on the reference length. Moreover, the harmonic mean values also show that M2 places more relevant content than M1 at the beginning of the text because limiting the number of sentences was more beneficial for M2 than for M1. Therefore, the additional data boosted the model in aspects that were not noticeable by using F1 measures.

## 7 Conclusions

We presented an approach for Radiology Report Summarization that continuously pre-trains a general domain BART model. This approach focuses on two main aspects: the use of biomedical data to adapt the model to this specific domain and the use of several pre-training tasks designed to inject linguistic knowledge and increase the abstractivity of the generated summaries. After the pre-training phase, we fine-tuned this model with different amounts of data from the shared task.

We also presented a study of the relationship between the models performance and the lengths of the generated summaries. We observed that our models condense the main information in the first sentences of the summaries. From the length distribution of the summaries, we found that our models tend to generate summaries with a common length; meanwhile, the reference summaries present more length variability. It seems that this behavior could penalize the performance of our

models, especially on those reports with short reference summaries.

## Limitations

The pre-training methodology used in this work applies a masking process at the sentence level that requires scoring the relevance of each sentence within the text. Therefore, this implies additional computational costs, limiting the scalability of our approach.

Due to time restrictions, the appearance of hallucinations in the generated radiology reports by our models has not been measured. It would be necessary to quantify this aspect because of the criticality of the domain of use in future works.

## Ethics Statement

The additional data that we have used for the pre-training process are from the MIMIC dataset, which meets the ethical requirements of Patient Health Information.

## Acknowledgements

This work is partially supported by MCIN/AEI/10.13039/501100011033, by the "European Union and "NextGenerationEU/MRR", and by "ERDF A way of making Europe" under grants PDC2021-120846-C44 and PID2021-126061OB-C41. It is also partially supported by the Generalitat Valenciana under project CIPROM/2021/023, and by the Spanish Ministerio de Universidades under the grant FPU21/05288 for university teacher training.

## References

Vicent Ahuir, Lluís-F. Hurtado, José Ángel González, and Encarna Segarra. 2021. Nasca and nases: Two monolingual pre-trained models for abstractive summarization in catalan and spanish. *Applied Sciences*, 11(21).

Asma Ben Abacha, Yassine Mrabet, Yuhao Zhang, Chaitanya Shivade, Curtis Langlotz, and Dina Demner-Fushman. 2021. Overview of the MEDIQA 2021 shared task on summarization in the medical domain. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 74–85, Online. Association for Computational Linguistics.

Asma Ben Abacha, Chaitanya Shivade, and Dina Demner-Fushman. 2019. Overview of the MEDIQA 2019 shared task on textual inference, question entailment and question answering. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 370–379, Florence, Italy. 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.

Songtai Dai, Quan Wang, Yajuan Lyu, and Yong Zhu. 2021. BDKG at MEDIQA 2021: System report for the radiology report summarization task. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 103–111, Online. Association for Computational Linguistics.

Jean-Benoit Delbrouck, Pierre Chambon, Christian Bluethgen, Emily Tsai, Omar Almusa, and Curtis Langlotz. 2022a. Improving the factual correctness of radiology report generation with semantic rewards. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4348–4360, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jean-benoit Delbrouck, Khaled Saab, Maya Varma, Sabri Eyuboglu, Pierre Chambon, Jared Dunnmon, Juan Zambrano, Akshay Chaudhari, and Curtis Langlotz. 2022b. ViLMedic: a framework for research at the intersection of vision and language in medical AI. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 23–34, Dublin, Ireland. Association for Computational Linguistics.

Jean-Benoit Delbrouck, Maya Varma, Pierre Chambon, and Curtis Langlotz. 2023. Overview of the radsum23 shared task on multi-modal and multi-anatomical radiology report summarization. In *Proceedings of the 22st Workshop on Biomedical Language Processing*, Toronto, Canada. Association for Computational Linguistics.

Wirth F Ferger. 1931. The nature and use of the harmonic mean. *Journal of the American Statistical Association*, 26(173):36–40.

A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. Ch. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley. 2000. PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals. *Circulation*, 101(23):e215–e220.

Max Grusky, Mor Naaman, and Yoav Artzi. 2018. NEWSROOM: A Dataset of 1.3 Million Summaries with Diverse Extractive Strategies. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Trans. Comput. Healthcare*, 3(1).

Karl Moritz Hermann, Tomás Kociský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *NIPS*, pages 1693–1701.

Alistair Johnson, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. 2023. MIMIC-IV-Note: Deidentified free-text clinical notes (version 2.2). PhysioNet.

Alistair Johnson, Tom Pollard, and Roger Mark. 2016. MIMIC-III Clinical Database (version 1.4). PhysioNet.

Alistair Johnson, Tom Pollard, Roger Mark, Seth Berkowitz, and Steven Horng. 2019. MIMIC-CXR Database (version 2.0.0). PhysioNet.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

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. Association for Computational Linguistics.

Richard Liaw, Eric Liang, Robert Nishihara, Philipp Moritz, Joseph E Gonzalez, and Ion Stoica. 2018. Tune: A research platform for distributed model selection and training. *arXiv preprint arXiv:1807.05118*.

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.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

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.

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, Rémi 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 M. 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.

Hongyi Yuan, Zheng Yuan, Ruyi Gan, Jiaxing Zhang, Yutao Xie, and Sheng Yu. 2022. BioBART: Pretraining and evaluation of a biomedical generative language model. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 97–109, Dublin, Ireland. Association for Computational Linguistics.

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

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020b. Bertscore: Evaluating text generation with bert.