Biomedical Data-to-Text Generation via Fine-Tuning Transformers

Ruslan Yermakov
Decision Science
& Advanced Analytics
Bayer AG
yermakovruslan@gmail.com

Nicholas Drago
Regulatory Policy
and Intelligence
Bayer AG

Angelo Ziletti*
Decision Science
& Advanced Analytics
Bayer AG
angelo.ziletti@bayer.com

Abstract

Data-to-text (D2T) generation in the biomedical domain is a promising - yet mostly unexplored - field of research. Here, we apply neural models for D2T generation to a real-world dataset consisting of package leaflets of European medicines. We show that fine-tuned transformers are able to generate realistic, multi-sentence text from data in the biomedical domain, yet have important limitations. We also release a new dataset (*BioLeaflets*) for benchmarking D2T generation models in the biomedical domain.

1 Introduction

Data-to-text (D2T) systems are attracting considerable interest due to their ability to automate the time-consuming writing of data-driven reports. There is a hitherto largely untapped potential for text generation in the biomedical domain. Potential applications of natural language generation of patient-friendly biomedical text include preparation of the first draft of package leaflets, patient education materials, or direct-to-consumer promotional materials in countries where this is permitted. Here we focus on a D2T task aiming to generate fluent and fact-based descriptions from biomedical data.

2 Related Work

Recently, neural D2T models have significantly improved the quality of short text generation (usually one sentence long) from input data compared to multi-stage pipelined or template-based approaches. Examples include biographies from Wikipedia fact tables (Lebret et al., 2016), restaurant descriptions from meaning representations (Novikova et al., 2017b), and basketball game summaries from statistical tables (Wiseman et al., 2017). Still, neural D2T approaches have major challenges, as outlined by Wiseman et al. (2017) and Parikh et al. (2020) which hinder their application to many real-world applications. These include hallucination effects (generated phrases not supported or contradictory to the source data), missing facts (generated text does not include input information), intersentence incoherence, and repetitiveness in the generated text. Following the success of leveraging pre-trained large-scale language models for a large variety of tasks, Kale and Rastogi (2020) fine-tuned T5 models (Raffel et al., 2020) for D2T generation. This strategy achieved state-of-the-art performance on task-oriented dialogue (MultiWoz) (Budzianowski et al., 2018), tables-to-text (ToTTo) (Parikh et al., 2020) and graph-to-text (WebNLG) (Gardent et al., 2017).

To the best of our knowledge, recent neural approaches and transfer learning strategies have not been applied to multi-sentence generation from input data, nor have they been applied in the biomedical domain. Our contribution is two-fold: we introduce a real-world biomedical dataset *BioLeaflets*, and demonstrate that transformers can generate high-quality multi-sentence text from data in the biomedical domain. The *BioLeaflets* dataset, fine-tuned models, code, and generated samples are available at https://github.com/bayer-science-for-a-better-life/data2text-bioleaflets.

3 The *BioLeaflets* Dataset

We introduce a new biomedical dataset for D2T generation - *BioLeaflets*, a corpus of 1336 package leaflets of medicines authorised in Europe, which we obtain by scraping the European Medicines Agency (EMA) website. This dataset comprises the large majority ($\sim$ 90%) of medicinal products authorised through the centralised procedure in Europe as of January 2021.

Package leaflets are published for medicinal products approved in the European Union (EU). They are included in the packaging of medicinal

Proceedings of the 14th International Conference on Natural Language Generation (INLG), pages 364–370,
Aberdeen, Scotland, UK, 20-24 September 2021.

(a) Original section content	**novonorm** is **an oral antidiabetic medicine** containing **repaglinide** which helps your **pancreas** produce more **insulin** and thereby lower **your blood sugar** (**glucose**). **type 2 diabetes** is **a disease** in which your **pancreas** does not make enough **insulin** to control **the sugar** in your blood or where your **body** does not respond normally to the **insulin** it produces. **novonorm** is used to control **type 2 diabetes** in adults as an add-on to diet and exercise: **treatment** is usually started if diet, **exercise and weight reduction** alone have not been able to control (or lower) **your blood sugar**. **novonorm** can also be given with **metformin**, another medicine for **diabetes**. **novonorm** has been shown to lower **the blood sugar**, which helps to prevent **complications** from **your diabetes**.
(b) Input: entities as a flat string	<PRODUCT_NAME> novonorm </PRODUCT_NAME> <TREATMENT> an_oral_antidiabetic_medicine </TREATMENT> <GENERIC_NAME> repaglinide </GENERIC_NAME> <SYSTEM_ORGAN_SITE> pancreas </SYSTEM_ORGAN_SITE> <GENERIC_NAME> insulin </GENERIC_NAME> <TEST> your_blood_sugar </TEST> ...<TEST> the_blood_sugar </TEST> <DX_NAME> complications </DX_NAME> <PROBLEM> your_diabetes </PROBLEM>
(c) T5 generated section	**novonorm** is **an oral antidiabetic medicine** containing **repaglinide** which helps your **pancreas** produce more **insulin** and thereby lower **your blood sugar** (**glucose**). **type 2 diabetes** is **a disease** in which your panccreases does not make enough **insulin** to control **the sugar** in your blood or where your **body** does not respond normally to the **insulin** it produces. novonourm is used to control type 2 diabetic patients who still have their **diabetes** and who are still unable to respond normally despite **treatment**, **exercise and weight reduction**, which have not helped to control your blood **glucose**. there is also a test that novonorgm can be given in combination with **metformin**, another medicine for **diabetes**. it is important that you continue to follow the advice on diet, exercising and diet that your nurse or doctor has given you.
(d) BART generated section	novondorm is **an oral antidiabetic medicine** containing **repaglinide** which helps your **pancreas** produce more **insulin** and thereby lower **your blood sugar** (**glucose**). **type 2 diabetes** is **a disease** in which your pancrea does not make enough **insulin** to control **the sugar** in your blood or where your **body** does not respond normally to the **insulin** it produces. **novonorm** is used to treat type 2 diabetic in adults. **treatment** is usually started with diet and **exercise and weight reduction**. your blood **glucose** may be increased when you start to take novonor on its own or in combination with **metformin**. if you have **diabetes**, no
(e) Content planner generated section	**novonorm** contains the active substance **repaglinide** which helps to lower **your blood sugar** (**glucose**). **type 2 diabetes** is **a disease** where your **body** does not make enough **insulin** to control **the sugar** in your blood or where your **body** does not respond normally to the **insulin** it produces. repaglinide krka is used to control **type 2 diabetes** in adults as **type 2 diabetes**. **type 2 diabetes** is also called non - insulin - dependent diabetes mellitus. **type 2 diabetes** is also a condition in which your **body** does not make enough **insulin** or the **insulin** that your **body** produces does not work as well as it should. your **body** can also make too much sugar. when this happens, sugar (**glucose**) builds up in the blood. this can lead to serious medical problems like heart disease, kidney disease, 2 and 2.

Table 1: Example of text generations. Entities are highlighted in bold, typos are underlined, and hallucinations are shown in red.

products and contain information to help patients use the product safely and appropriately, under the guidance of their healthcare professional. Package leaflets are required to be written in a way that is clear and understandable (EU, 2001). Each document contains six sections (see Table 2).The main challenges of this dataset for D2T generation are multi-sentence and multi-section target text, small sample size, specialized medical vocabulary and syntax.

3.1 Dataset Construction

The content of each section is not standardized, yet it is still well-structured. Thus, we identify sections via heuristics such as regular expressions and word overlap. The content of each section is lower-cased and tokenized by treating all special characters as separate tokens. Duplicates are also removed. We randomly split the dataset into training (80%), development (10%), and test (10%) set. Table 2 summarizes dataset statistics.

3.2 Dataset Annotations

We do not have annotations available for the package leaflet text. To create the required input for D2T generation, we augment each document by leveraging named entity recognition (NER). Parikh et al. (2020) indicated it is important that target summaries contain information that can be inferred from the input data to avoid dataset-induced hallucinations. To this end, we combine two NER frameworks: Amazon Comprehend Medical (ACM) (Bhatia et al., 2019) and Stanford Stanza (Qi et al., 2020; Zhang et al., 2021). ACM and Stanza achieved entity micro-averaged test F1 of 85.5% and 88.13% respectively on the 2010 i2b2/VA clinical dataset (Uzuner et al., 2011). We further leverage ACM to detect medical conditions from ICD-10 (WHO, 2004) and medications from RxNorm. Additionally, we treat all digits as entities, and add the medicine name as first entity. In case of overlapping entities from different sources, we favor longer entities over shorter ones. As a result of the NER

Section type	No. samples	Average length (characters)	Average length (tokens)	Average no. entities per section	No. unique entities
1. What the product is and what it is used for	1 314	963	174	29.3	9 641
2. What you need to know before you take the product	1 309	4 560	849	127.7	23 278
3. How to take the product	1 313	2 300	458	50.5	11 640
4. Possible side effects	1 295	3 453	651	135.2	27 945
5. How to store the product	1 172	631	123	6.3	2 041
6. Content of the pack and other information	1 311	982	196	38.4	9 932

Table 2: *BioLeaflets* dataset statistics grouped by section type.

process, we obtain 26 unique entity types. Examples are: *problem*: ('active chronic hepatitis', 'migraine pain'), *system-organ-site*: ('blood vessel', 'kidneys', 'surrounding tissue'), *treatment*: ('routine dental care', 'a vaccination', 'a chemotherapy medicinal product'), or *procedure*: ('injections', 'spinal or epidural anaesthesia', 'surgical intervention', 'bone marrow or stem cell transplant').

BioLeaflets proposes a conditional generation task: given an ordered set of entities as source, the goal is to produce a multi-sentence section. Since only the entities are provided as input, the structured data is underspecified. A human without specialized knowledge would likely be unable to produce satisfactory text. However, we expect that a labeling expert with profound knowledge of package leaflets would be able to generate (with some difficulty) satisfactory text in the large majority of cases. Successful generation thus requires the model to learn specific syntax, terminology, and writing style from the corpus (e.g., via fine-tuning).

4 Experiments

Following Kale and Rastogi (2020), we represent the structured data (i.e., detected entities) as a flat string (linearization). The entities are kept in their order of appearance (Table1b). The models are then trained to predict - starting from these entities - the corresponding published leaflet text.

We present baseline results on *BioLeaflets* dataset by employing the following state-of-the-art approaches:

- **Content Planner**: two stages neural architecture (content selection and planning) based on LSTM (Puduppully et al., 2019). Since only relevant entities are provided as input to the model, we solely use the content planning stage (encoder-decoder architecture with an attention mechanism). We train one model for each section, and use the same hyperparameters reported by Puduppully et al. (2019).

- **T5**: a text-to-text transfer transformer model (Raffel et al., 2020). Kale and Rastogi (2020): showed that T5 outperforms alternatives like BERT (Devlin et al., 2019) and GPT-2 (Radford et al., 2019). After hyperparameter search on the development dataset, the following parameters (yielding the best ROUGE-L score (Lin, 2004)) are selected: constant learning rate of 0.001, batch size of 32, 20 epochs, greedy search as a decoding method.

- **BART**: denoising autoencoder for pretraining sequence-to-sequence models with transformers (Lewis et al., 2020). For computational reasons, we use the same hyperparameters as per T5 fine-tuning.

- **BART and T5 with conditioning**: we add the prefix "section_n" ($n = 1, \ldots 6$) to the (linearized) input data. This explicitly gives the model information on the section number and thus enforces a conditioning on the section type for text generation.

BART and T5 fine-tuning are performed via HuggingFace (Wolf et al., 2020).

5 Evaluation

Table 1 shows the generated text for one test sample as illustrative example. All generated text is

Model	Word-overlap metrics		Semantic equivalence metrics		
	SacreBLEU	ROUGE-L	BERTScore	BLEURT	MoverScore-2l
Content Planner	**27.78**	39.32	0.214	-0.072	0.591
BART-base	8.76 ± 0.02	42.73 ± 0.11	**0.370** ± 0.001	**0.268** ± 0.002	0.609 ± 0.0004
BART-base + cond	8.73 ± 0.02	42.60 ± 0.12	**0.369** ± 0.001	**0.268** ± 0.003	0.608 ± 0.0004
T5-base	18.68 ± 0.07	**47.22** ± 0.17	0.363 ± 0.001	0.255 ± 0.008	**0.620** ± 0.0005
T5-base + cond	18.63 ± 0.14	**47.31** ± 0.22	0.364 ± 0.002	0.256 ± 0.006	**0.621** ± 0.0008

Table 3: Results on the BioLeaflets test set (averaged over all sections). T5 and BART models are fine-tuned with seven different random seeds: average and standard deviation are reported. BLEURT-large-128 is used.

Model		Adequacy	Hallucination presence	Entity inclusion	Fluency
Content Planner	annotator 1	4.1 ± 3.0	6.8 ± 3.2	4.8 ± 3.2	5.1 ± 3.3
	annotator 2	3.7 ± 2.6	6.4 ± 2.5	5.1 ± 2.5	5.4 ± 2.3
BART-base	annotator 1	7.5 ± 2.1	3.1 ± 2.6	7.4 ± 2.3	8.6 ± 1.8
	annotator 2	**6.6** ± 2.2	**3.3** ± 2.1	**8.1** ± 1.8	8.0 ± 1.3
T5-base	annotator 1	**7.8** ± 1.8	**3.0** ± 2.4	**7.6** ± 2.1	**9.0** ± 1.4
	annotator 2	6.5 ± 2.2	3.5 ± 1.9	7.8 ± 1.7	**8.2** ± 1.2

Table 4: Human evaluation of test samples. Values on a scale from one to ten; average and standard deviation are reported. The higher the better for all quantities, expect for "Hallucination presence". Adequacy estimates the overall generation quality, taking into consideration fluency, amount of hallucination, and entities included in the generated text.

made available[1]. After a thorough inspection of the samples, we conclude that generated text is generally fluent and coherent. Text produced by T5 and BART is more fluent, factually and grammatically correct than those by Content Planner. Table 3 illustrates the performance of state-of-the-art models quantified by automatic metrics. Word-overlap metrics such as (Sacre)BLUE (Post, 2018) and ROUGE (Lin, 2004) have been shown to perform poorly in evaluation of natural language generation (Novikova et al., 2017a), and thus we report them here only for completeness. Conversely, contextual embedding based metrics BERTScore (Zhang* et al., 2020), BLEURT (Sellam et al., 2020), and MoverScore-2 (Zhao et al., 2019) correlate with human judgment on sentence-level and system-level evaluation. They adequately capture semantic equivalence between generated and target text as well as fluency and overall quality. T5 and BART outperform Content Planner, as measured by BERTscore, BLEURT, and MoverScore-2. T5 and BART show similar performance. These results show that transformer-based models and transfer learning strategies achieve state-of-the-art performance on data-to-text tasks, generalizing the findings in Kale and Rastogi (2020) to multi-sentence and multi-section generation, biomedical text, and low-data setting.

To confirm these findings, human evaluation is performed for Section 1 of the test set by two annotators. Results are shown in Table 4. Similarly to Manning et al. (2020), we design a survey which includes adequacy (estimate of overall quality), presence of hallucinations, entity inclusion, and fluency. T5 and BART have similar performance, and they produce more adequate text than Content Planner. T5 and BART performance is more stable across samples (lower standard deviation). These conclusions coincide with the ones drawn from Table 3, thus confirming the usefulness of semantic equivalence metrics for automatic evaluation of text generation.

Interestingly, specifying the section type in the input records (i.e., explicit conditioning) did not improve model performances (Table 3). To rationalize this result, we analyze T5 internal representations. Specifically, for each test sample, we extract the (average) last encoder hidden-state for both pre-trained (not fine-tuned) and fine-tuned T5 (fine-tuned on *BioLeaflets* but without explicit

[1] https://github.com/bayer-science-for-a-better-life/data2text-bioleaflets

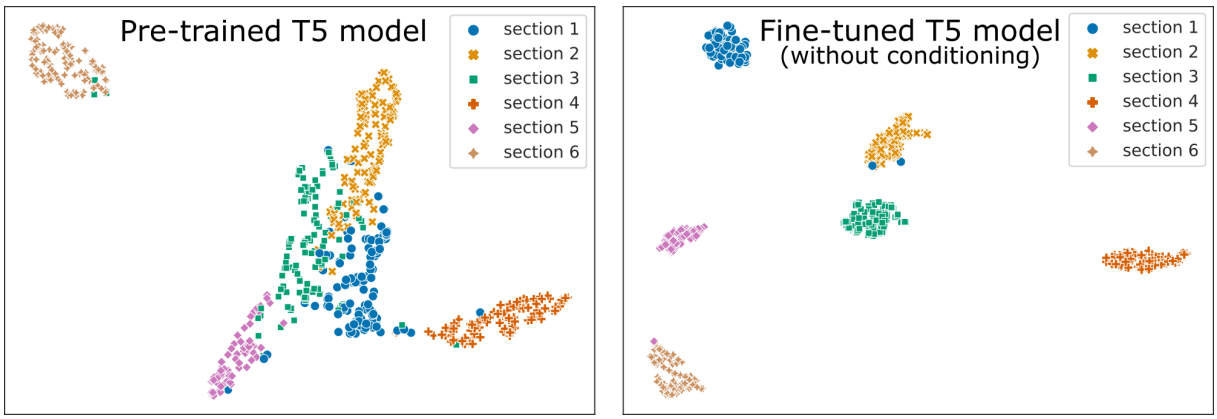

Figure 1: Two-dimensional projections of T5 internal representations (average of the last encoder hidden-states) for pre-trained (not fine-tuned) (**left**) and fine-tuned T5 model on *BioLeaflets* dataset (**right**). T5 implicitly learns to condition on section type during fine-tuning.

conditioning). We then project these vectors into two-dimensions using the non-linear dimensionality reduction method UMAP (McInnes et al., 2020). The results are depicted in Fig. 1. In Fig. 1 (right), we can identify six well-separated clusters, which correspond to (the internal representations of samples belonging to) the six document sections in the *BioLeaflets* dataset. Thus, after fine-tuning, T5 maps input data belonging to different sections to different parts of the internal representation space. The cluster separation is much less pronounced for the pre-trained (not-fine-tuned) T5 model (Fig. 1, left). This shows that during the fine-tuning process, T5 implicitly learns to condition on section type, thus learning to generate different sections, even despite the small dataset. Since conditioning is learned automatically, explicitly passing the section type as input does not increase model performance.

6 Error Analysis and Limitations

After thorough qualitative evaluation of numerous generated samples, the following general issues appear:

- **Typos**: Even though models largely utilize the input entities correctly, typos appear in generated text by T5 and BART for out-of-vocabulary words, e.g. Table 1 (c, d). Content Planner does not seem to have this problem.

- **Hallucinations** are present for all models. Loss functions like maximum likelihood do not directly minimize hallucinations, thus hindering consistent fact-based text generation.

- **Repetitiveness**: Content Planner produce repetitions (e.g. Table 1 (e)), whereas T5 and BART language models do not.

- **Difficulties in producing coherent long text**: In the *BioLeaflets* dataset, models perform well in generating section 1, which is 962 characters long on average. However, the quality of section 4 "Possible side effects" (3 453 characters long on average) generation is poor.

Possible improvements to our work are: analysis of the impact of shuffling of entities for the input data generation, introduction of loss functions that explicitly favor factual correctness, usage of specialized biomedical embeddings, inclusion of more source input data (e.g. part-of-speech, dependency tag), generation of longer text (beyond the 512 tokens generated here).

7 Conclusion

In this study, we introduce a new biomedical dataset (*BioLeaflets*), which could serve as a benchmark for biomedical text generation models. We demonstrate the feasibility of generating coherent multi-sentence biomedical text using patient-friendly language, based on input consisting of biomedical entities. These results show the potential of text generation for real-world biomedical applications. Nevertheless, human evaluation is still a required step to validate the generated samples. Application of the methodology and models used here to different sets of biomedical text (e.g., generation of selected sections of clinical study reports) could be an area for further research.

References

Parminder Bhatia, Busra Celikkaya, Mohammed Khalilia, and Selvan Senthivel. 2019. Comprehend medical: A named entity recognition and relationship extraction web service. In *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*, pages 1844–1851.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. MultiWOZ - a large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5016–5026, Brussels, Belgium. 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.

European Union EU. 2001. Directive 2001/83/ec of the european parliament and of the council of 6 november 2001 on the community code relating to medicinal products for human use. Brussels, Belgium.

Claire Gardent, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. 2017. The WebNLG challenge: Generating text from RDF data. In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 124–133, Santiago de Compostela, Spain. Association for Computational Linguistics.

Mihir Kale and Abhinav Rastogi. 2020. Text-to-text pre-training for data-to-text tasks. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 97–102, Dublin, Ireland. Association for Computational Linguistics.

Rémi Lebret, David Grangier, and Michael Auli. 2016. Neural text generation from structured data with application to the biography domain. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1203–1213, Austin, Texas. Association for Computational Linguistics.

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.

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.

Emma Manning, Shira Wein, and Nathan Schneider. 2020. A human evaluation of amr-to-english generation systems. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 4773–4786. International Committee on Computational Linguistics.

Leland McInnes, John Healy, and James Melville. 2020. Umap: Uniform manifold approximation and projection for dimension reduction.

Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. 2017a. Why we need new evaluation metrics for NLG. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2241–2252, Copenhagen, Denmark. Association for Computational Linguistics.

Jekaterina Novikova, Ondřej Dušek, and Verena Rieser. 2017b. The E2E dataset: New challenges for end-to-end generation. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 201–206, Saarbrücken, Germany. Association for Computational Linguistics.

Ankur P Parikh, Xuezhi Wang, Sebastian Gehrmann, Manaal Faruqui, Bhuwan Dhingra, Diyi Yang, and Dipanjan Das. 2020. ToTTo: A controlled table-to-text generation dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 1173–1186. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Ratish Puduppully, Li Dong, and Mirella Lapata. 2019. Data-to-text generation with content selection and planning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):6908–6915.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

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

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.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Özlem Uzuner, Brett R South, Shuying Shen, and Scott L DuVall. 2011. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5):552–556.

WHO WHO. 2004. Icd-10 : international statistical classification of diseases and related health problems : tenth revision.

Sam Wiseman, Stuart Shieber, and Alexander Rush. 2017. Challenges in data-to-document generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2253–2263, Copenhagen, Denmark. Association for Computational Linguistics.

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.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Yuhao Zhang, Yuhui Zhang, Peng Qi, Christopher D Manning, and Curtis P Langlotz. 2021. Biomedical and clinical English model packages for the Stanza Python NLP library. *Journal of the American Medical Informatics Association*.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 563–578, Hong Kong, China. Association for Computational Linguistics.