Stylistic MR-to-Text Generation Using Pre-trained Language Models

Kunal Pagarey, Kanika Kalra, Abhay Garg, Saumajit Saha, Mayur Patidar, Shirish Karande
TCS Research Pune, India
{kunal.pagarey, kalra.kanika, abhay.garg1, saha.saumajit,
patidar.mayur, shirish.karande}@tcs.com

Abstract

We explore the ability of pre-trained language models BART, an encoder-decoder model, GPT2 and GPT-Neo, both decoder-only models for generating sentences from structured MR tags as input. We observe best results on several metrics for the YelpNLG and E2E datasets. Style based implicit tags such as emotion, sentiment, length etc., allows for controlled generation but it is typically not present in MR. We present an analysis on YelpNLG showing BART can express the content with stylistic variations in the structure of the sentence. Motivated with the results, we define a new task of emotional situation generation from various POS tags and emotion label values as MR using EmpatheticDialogues dataset and report a baseline. Encoder-Decoder attention analysis shows that BART learns different aspects in MR at various layers and heads.

1 Introduction

Recent advances in NLG focus on generating text from structured data encoded as Meaningful Representations (MR). MR typically comprises of semantic content to be realized for generation. This can be used for automating writing reports from tabular data, descriptions and reviews for products or restaurants from catalog, etc. However, style based implicit tags can add dynamic, engaging and immersive effect in real world NLG applications such as social and empathetic chatbots. The style aspects along with content information allows generating varied and customized text with same content. In this work, we explore capabilities of an encoder-decoder model, BART (Lewis et al., 2019), and two decoder-only models, GPT2 (Radford et al., 2019) and GPT-Neo (Black et al., 2021) for MR-to-text generation task. We evaluate BART, GPT2 and GPT-Neo on three datasets, one for content and other for both content and style. These datasets include E2E original and clean version (Dušek et al., 2020) (Dušek et al., 2019) which are restaurant description datasets comprising of content based MR and Yelp NLG (Oraby et al., 2019), a restaurant reviews corpus having both semantic and stylistic tags. We define a new task of emotional situation generation on Empathetic dialogues (ED) dataset (Rashkin et al., 2018). We construct MRs using set of POS tag (Qi et al., 2020) values from situation along with emotion label. Table 9 of Appendix A describes sample input MR and output for each dataset.

Our main contributions are defined as: a) The ability of encoder-decoder based and decoder-only pretrained transformer models to generate fluent sentences from content and style based MR. b) A new task on emotional situation generation using POS tag and emotion label values as MR and report its baseline. c) Encoder-Decoder attention map analysis of BART to further understand which layer and head learns which concept.

2 Related Work

Existing structured data to text datasets - E2E (Dušek et al., 2020) (Dušek et al., 2019), WebNLG (Gardent et al., 2017), TOTTO (Parikh et al., 2020), AGENDA (Koncel-Kedziorski et al., 2019) etc consider input in various formats such as slot value pair, triplets, or graph. They consist of content based semantic input in MR. Recently introduced YelpNLG dataset by (Oraby et al., 2019) considers style aspect in addition to content slot value in MR and provides LSTM encoder decoder baseline. Our work focuses on exploring recent language model capability for content and style based MR.

Researchers have attempted to improve content slot value MR to text in attention based encoder decoder architectures by incorporating various techniques. (Tseng et al., 2020) performed joint training of NLU and NLG. (Roberti et al., 2019) in-

Proceedings of the 18th International Conference on Natural Language Processing, pages 93–99
Silchar, India. December 16 - 19, 2021.

Dataset	Size	Content	Style
Yelp	300k	restaurant[],cuisine[], food[], staff[], service[], ambiance[], price[]	Sentiment(positive, negative, neutral), length(short,medium,long), perspective(first,person, not first person),exclamation (has exclamation, no exclamation)
E2E	50k	name[], eatType[],food[],near[], priceRange[], customerRating[], area[], kidsFriendly[]	NA
ED	25k	POS values subset from [Noun], [Adjective], [Verb], [Pronoun]	32 Emotion Labels

Table 1: Content and style tag description for each dataset. ED only consists of values without content slot type.

troduced copy mechanism from MR facts to text. (Kedzie and McKeown, 2020) performed controllable MR-to-text generation by comparing different linearization strategies and phrase-based data augmentation technique. (Juraska et al., 2018), (Zhang et al., 2018), (Gong, 2018) applied reranking on top of seq2seq model providing semantic control, (Puzikov and Gurevych, 2018) came up with data-driven and template-based generation system. (Shen et al., 2019) used computational pragmatic based approach for conditional generation. However, we observe that all pre-trained transformer models perform well irrespective of their sizes, without requiring changes for both content and style MR.

3 Dataset Description

Table 1 provides a concise description of the datasets used, which were constructed to explore and improve the natural language generation capability of neural architectures. E2E (Dušek et al., 2020) original, a restaurant review dataset, has high lexical diversity and diverse discourse phenomena. E2E clean by (Dušek et al., 2019) is a noise free version of E2E (Dušek et al., 2020), with no mismatch between the content of the MR tags and the corresponding references. (Oraby et al., 2019) curated MR for YelpNLG automatically by leveraging freely available user review data on restaurants. This dataset brings in rich language descriptions with varied semantic emotions and content. To further explore the empathetic conversational potential, we use ED dataset (Rashkin et al., 2018), which comprises emotional dialogues between two persons. Motivated by YelpNLG, we constructed MR using values from POS tag set from noun, adj, pronoun and emotion label values for emotional situations provided in ED dataset.

4 Experiments

We fine-tune pre-trained language models like BART-large, GPT2-medium and GPT-Neo 125M for MR-to-text on train split of respective datasets. We use early stopping and choose the best model for evaluation on test set. Other parameters used for fine tuning are AdamW optimizer with a learning rate of 3e-5 and a linear learning rate scheduler. While generating the output text from MR we use beam search decoding with beam size of 4. We evaluate the generated output text using the standard automatic evaluation metrics[1] BLEU (Papineni et al., 2002), METEOR (Lavie and Agarwal, 2007), NIST (Doddington, 2002), CIDEr (Vedantam et al., 2015) and ROUGUE (Lin, 2004), and Semantic Error Rate (SER) (Dušek et al., 2019).

5 Results and Discussion

E2E, E2E clean: We report the fine-tuning results of the pre-trained models for E2E in Table 2 and compare it with other recent baselines[2]. We obtain best METEOR, CIDEr and ROUGE-L scores for E2E original using GPT2 and for E2E clean, best NIST using GPT2 and best SER score using BART. The other scores are comparable with baselines and do not differ significantly. The results show that the pre-trained models are able to preserve the content tags in output.

YelpNLG: We report the results for Yelp NLG in Table 3. We consider all the settings for YelpNLG - only content (BASE), content with style addition at different granularity (adjectives, sentiment, all other style aspects) . We obtain best results on all the metrics excluding SER using BART. SER less than 5% for BASE and STYLE setting signifies that

[1] https://github.com/tuetschek/e2e-metrics

[2] We show few baseline scores due to space constraint.

	Architectures	BL(↑)	NT(↑)	MT(↑)	RL(↑)	CD(↑)	SER(↓)
E2E Original	(Dušek and Jurčíček, 2016)	0.6593	8.6094	0.4483	0.685	2.2338	3.56*
	(Zhang et al., 2018)	0. 6545	8.184	0.4392	0.7083	2.1012	-
	(Tseng et al., 2020)	0.6855	-	-	-	-	-
	(Shen et al., 2019)	**0.6860**	**8.73**	0.4525	0.7082	2.37	-
	BART	0.6757	8.7242	0.4614	0.703	2.3914	3.58
	GPT2	0.6853	8.7164	**0.4637**	**0.7143**	**2.411**	5.56
	GPT-Neo	0.6841	8.6654	0.4626	0.7064	2.3697	**3.52**
E2E Clean	(Dušek and Jurčíček, 2016)	0.4073	6.1711	0.3776	0.5609	1.8518	0.87
	(Harkous et al., 2020)	**0.436**	-	**0.39**	**0.575**	**2.0**	-
	BART	0.4258	6.4188	0.3858	0.5677	1.9355	**0.13**
	GPT2	0.4285	**6.4524**	0.3854	0.5718	1.9873	1.02
	GPT-Neo	0.4087	6.2472	0.3751	0.5561	1.7928	4.06

Table 2: Results on E2E original & Clean test set. * - score on provided outputs. All tables follow these abbreviations - BL: BLEU, NT: NIST, MT: METEOR, RL: Rouge-L, CD: CIDEr

	Variant	BL	MT	CD	NT	SER
Baseline	Base	0.126	0.206	1.300	3.840	0.053
	+Adj	0.164	0.233	1.686	4.547	0.063
	+Sent	0.166	0.234	1.692	4.477	0.064
	+Style	0.173	0.235	1.838	5.537	0.090
Bart	Base	**0.177**	**0.227**	**1.820**	**5.303**	0.0346
	+Adj	**0.224**	**0.263**	**2.355**	**6.130**	0.0358
	+Sent	**0.225**	**0.264**	**2.358**	**6.158**	0.0382
	+Style	0.226	**0.268**	**2.587**	6.143	0.0435
Gpt2	Base	0.1673	0.2235	1.7731	4.7605	**0.0291**
	+Adj	0.2057	0.2578	2.2868	4.8509	**0.0308**
	+Sent	0.2072	0.2594	2.2971	4.8365	**0.0302**
	+Style	**0.2276**	0.2648	2.5799	**6.3915**	**0.0337**
GptNeo	Base	0.1646	0.2181	1.6502	5.052	0.0345
	+Adj	0.1972	0.2546	2.2095	4.6360	0.0320
	+Sent	0.2006	0.2548	2.2104	4.8331	0.0315
	+Style	0.2223	0.2611	2.5034	6.3503	0.0443

Table 3: Results on YelpNLG test set. Base MR only contains content slot type-value pairs, +Adj contains content slot type-value-adjective triplets. In addition to +Adj, sentiment and other stylistic aspects are added in +Sent and +Style, respectively.

	Variant	BL	MT	CD	NT	RL
Bart	NAd	**0.245**	**0.293**	**2.414**	**6.939**	**0.539**
	NAdP	**0.358**	**0.349**	**3.638**	**8.383**	**0.660**
Gpt2	NAd	0.1855	0.2589	1.8918	5.9274	0.4852
	NAdP	0.2726	0.3071	2.8072	7.2803	0.5994
GptNeo	NAd	0.1389	0.2392	1.6136	4.6187	0.4474
	NAdP	0.2263	0.2925	2.4268	6.8125	0.5829

Table 4: Results on ED test set. NAd: noun+adjective, NAdP: NAd+pronoun

the pre-trained models understand and express the content specifications well in the generated review. The models learn to associate the attribute values in the MR tag even in presence of different stylistic aspects in the fairly complex sentences.

ED: We report the results for ED dataset in Table 4. For ED, we provide baseline for various sequences of POS tag values in MR - (Noun,Adj), (Noun,Adj,Pronoun), with emotion label. We also find here that BART performs best compared to other pretrained models. We observe that increasing the content value information leads to increment in scores. We observe that the emotional aspect of the generated statements can be manipulated by changing the input tags, which emphasizes the models' power to generate customized sentences while expressing all the relevant content as shown in Table 5 and Table 6.

5.1 Attention Map Analysis for YelpNLG

To further investigate the rationale behind the superior performance of the transformer models, we extend (Vig, 2019) to analyse the encoder decoder attentions of BART and report results in Table 7. As in CNNs, we hypothesised that different nodes in the architecture would learn to capture particular tags in the MR. Thus, we evaluate which head and layer learns to attend to a particular style aspect. LxHy signifies encoder-decoder attention map corresponding to yth head of xth layer. We take 1000 samples in each case and perform element wise summation of attention distribution scores of the chosen output tokens shown under 'By' over all input tokens. We report the percentage of cases in which the input token under 'To' was in top3

YelpNLG MR	Output
food||chicken||delicious||mention_1 food||fries||no_adj||mention_1,positive,len_medium,not_first_person,no_exclamation	the chicken was delicious and the fries were cooked to perfection.
food||chicken||delicious||mention_1 food||fries||no_adj||mention_1,positive,len_medium,not_first_person,has_exclamation	the chicken is delicious and the fries are to die for!
food||chicken||tasty||mention_1 food||fries||delicious||mention_1,negative,len_long,not_first_person,has_exclamation	the chicken was tasty, the fries were delicious, but the whole thing was a let down!
food||fries||tasteless||mention_1 food||chicken||yuck||mention_1,positive,len_medium,first_person,has_exclamation	the fries were tasteless, the chicken was yuck, but i love this place!

Table 5: Generated outputs for YelpNLG using BART.

ED	Output
joyful(friends different country)	My friends are coming to visit me from a different country.
guilty(friends different country)	I lied to my friends that I was going to visit them in a different country.
anxious(husband end month little worried funds)	My husband and I are going overseas at the end of the month. I am a little worried about how we will manage the funds.
Confident(husband end month little worried funds)	My husband and I are going to get married at the end of the month. I'm a little worried about the funds we'll have, but I know we'll make it happen.

Table 6: Generated outputs for ED for Emotion with Noun and Adjective using BART.

(top5 for 'all') most maximally attended input tokens. 'Sample Type' column denotes the common stylistic aspect for all those samples. We observe that various layers and heads learn different stylistic concepts beyond just learning to copy (as shown in Figure 1, 2, 3 and 4 of Appendix C). The results reinforce our hy pothesis and establish that different parts of BART learn to understand the intrinsic meaning of different tags.

5.2 Qualitative Analysis

The generated outputs emphasize sensitivity of BART towards stylistic aspects. A minute change from no_exclamation in first row to has_exclamation in second row in Table 5, to our surprise, BART has generated very different and dramatic output. The last two rows show BART's capability of handling contrasting scenarios wherein the sentiment of the input is in contrast to the adjective values of food. Results in Table 6 show that BART can express the same content with different emotions fed as implicit tags.

While analysing the predictions of the finetuned models, we observe that most of the time, BART has been successful in generating output as per the given style constraints mentioned in the input MR tags. Table 8 shows a few instances where BART, out of all the 3 models, is capable of producing better and more meaningful sentences. However in the last two examples of Table 8, we also report cases where these models produce incoherent outputs that do not meet the required stylistic aspects.

LH	Token Maximally Attended (To,By)	%	Sample Type
L7H0	('negative', all)	86	negative
L7H0	('neutral', all)	60.5	neutral
L1H11	('first', first person words)	100	first person
L3H8	('short', all)	99.4	short
L8H5	('has', '!')	87.7	has exclamation
L7H11	(adjectives, food values)	73	adjs
L0H15	(food values, food values)	100	food values

Table 7: Attention Analysis for YelpNLG.

5.3 Analysis of the Pre-trained models

We have used pretrained models like BART-large, GPT2-medium and GPTNeo-125M in our experiments. The number of parameters used for these models vary. For example, BART-large has 406 million parameters, GPT2-medium has 345 million parameters and GPTNeo-125M has only 125 million parameters. Table 2, 3, 4 show that GPTNeo-125M, inspite of having less than 50% of the param-

MR: food||steak||awesome||mention_1 food||chicken||fried||mention_1,neutral,len_long,first_person, no_exclamation **Ref:** "what i had in mind, try out their awesome steak and their fried chicken... however, it turns out they are only served on weekdays." **Bart:** "the steak was awesome, but the fried chicken was a bit too sweet for my taste buds, but it was still good." **Gpt2:** "i had the steak which was awesome, and my friend had the fried chicken which was also good." **GptNeo:** "the steak was awesome, but the fried chicken was a little dry for my taste."
MR: food||wontons||no_adj||mention_1 food||wontons||no_adj||mention_2 food||meat||no_adj||mention_1,positive,len_medium,not_first_person,no_exclamation **Ref:** "if you go anywhere in california, wontons are wontons with meat inside." **Bart:** "the wontons are not just wontons, they are filled with meat and veggies." **Gpt2:** "wontons are wontons, meat is meat." **GptNeo:** "wontons, fried wontons, fried wontons, fried wontons, fried wontons, fried wontons, fried wontons, fried wontons, fried wontons."
MR: food||chicken||lemon||mention_1 food||vegetables||no_adj||mention_1,negative,len_long,not_first_person,no_exclamation **Ref:** "first the appetizer was not prepared properly, out of the lemon chicken and then vegetables were cold." **Bart:** "the lemon chicken had a lot of vegetables in it, but not enough of them to make up for the lack of taste." **Gpt2:** "the lemon chicken had a lot of vegetables in it, but it was very dry." **GptNeo:** "the lemon chicken had a lot of vegetables in it, but it was a little dry."
MR: food||chicken||no_adj||mention_1 food||sauce||no_adj||mention_1,negative,len_medium,first_person,no_exclamation **Ref:** i did not like the way the chicken was cooked or the sauce on it. **Bart:** i had the chicken with the sauce on the side and it was good. **Gpt2:** i had the chicken with the sauce on the side. **GptNeo:** i had the chicken with the sauce on the side.
MR: food||pizza||no_adj||mention_1 food||steak||no_adj||mention_1 food||cheese||no_adj||mention_1,positive,len_medium,not_first_person,no_exclamation **Ref:** love the pizza the steak and cheese and the place. **Bart:** "pizza, wings, steak, cheese, and more." **Gpt2:** "pizza, steak, cheese, etc." **GptNeo:** "pizza, steak, and mac and cheese."

Table 8: Example outputs for YelpNLG using BART, GPT2 and GPTNeo.

eters of the other models, is being able to produce competitive results on these tasks, often beating the baseline performances in several cases.

GPTNeo has been pretrained on Pile (Gao et al., 2020) dataset, which is composed of different constituent sub-datasets from diverse domains. However, GPT2 and BART are pretrained exclusively on text data. The size of the pre-training dataset seems to have an impact in the performance of the pre-trained model on downstream tasks. This is because GPTNeo is trained on 800GB Pile dataset while GPT2 has been trained on only 40GB of webtext data.

6 Conclusion

We describe the benefits and importance of MR2text generation. We fine-tune recently introduced Transformer-based language models like BART, GPT2 and GPTNeo, and produce results on two versions of E2E, YelpNLG and ED datasets. We have defined a new task on Emphatetic Dataset to emphasize the usefulness of implicit tags in NLG. Quantitative and Qualitative analyses show how well BART captures the specifications and brings stylistic variations in generated outputs.

References

Sid Black, Leo Gao, Phil Wang, Connor Leahy, and Stella Biderman. 2021. GPT-Neo: Large scale autoregressive language modeling with mesh-tensorflow.

George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the second international conference on Human Language Technology Research*, pages 138–145.

Ondřej Dušek, David M. Howcroft, and Verena Rieser. 2019. Semantic noise matters for neural natural language generation. In *Proc. of the 12th International Conference on Natural Language Generation*, pages 421–426, Tokyo, Japan. Association for Computational Linguistics.

Ondřej Dušek and Filip Jurčíček. 2016. Sequence-to-sequence generation for spoken dialogue via deep syntax trees and strings. *arXiv preprint arXiv:1606.05491*.

Ondřej Dušek, Jekaterina Novikova, and Verena Rieser. 2020. Evaluating the State-of-the-Art of End-to-End Natural Language Generation: The E2E NLG Challenge. *Computer Speech & Language*, 59:123–156.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. 2020. The Pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.

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.

Heng Gong. 2018. Technical report for e2e nlg challenge. *E2E NLG Challenge System Descriptions*.

Hamza Harkous, Isabel Groves, and Amir Saffari. 2020. Have your text and use it too! end-to-end neural data-to-text generation with semantic fidelity. *arXiv preprint arXiv:2004.06577*.

Juraj Juraska, Panagiotis Karagiannis, Kevin K Bowden, and Marilyn A Walker. 2018. A deep ensemble model with slot alignment for sequence-to-sequence natural language generation. *arXiv preprint arXiv:1805.06553*.

Chris Kedzie and Kathleen McKeown. 2020. Controllable meaning representation to text generation: Linearization and data augmentation strategies. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5160–5185, Online. Association for Computational Linguistics.

Rik Koncel-Kedziorski, Dhanush Bekal, Yi Luan, Mirella Lapata, and Hannaneh Hajishirzi. 2019. Text generation from knowledge graphs with graph transformers. *arXiv preprint arXiv:1904.02342*.

Alon Lavie and Abhaya Agarwal. 2007. Meteor: An automatic metric for mt evaluation with high levels of correlation with human judgments. In *Proceedings of the second workshop on statistical machine translation*, pages 228–231.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Shereen Oraby, Vrindavan Harrison, Abteen Ebrahimi, and Marilyn Walker. 2019. Curate and generate: A corpus and method for joint control of semantics and style in neural nlg. *arXiv preprint arXiv:1906.01334*.

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.

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. *arXiv preprint arXiv:2004.14373*.

Yevgeniy Puzikov and Iryna Gurevych. 2018. E2e nlg challenge: Neural models vs. templates. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 463–471.

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*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2018. Towards empathetic open-domain conversation models: A new benchmark and dataset. *arXiv preprint arXiv:1811.00207*.

Marco Roberti, Giovanni Bonetta, Rossella Cancelliere, and Patrick Gallinari. 2019. Copy mechanism and tailored training for character-based data-to-text generation. In *Joint European Conference*

on Machine Learning and Knowledge Discovery in Databases, pages 648–664. Springer.

Sheng Shen, Daniel Fried, Jacob Andreas, and Dan Klein. 2019. Pragmatically informative text generation. *arXiv preprint arXiv:1904.01301*.

Bo-Hsiang Tseng, Jianpeng Cheng, Yimai Fang, and David Vandyke. 2020. A generative model for joint natural language understanding and generation. *arXiv preprint arXiv:2006.07499*.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.

Jesse Vig. 2019. A multiscale visualization of attention in the transformer model. *arXiv preprint arXiv:1906.05714*.

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. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Biao Zhang, Jing Yang, Qian Lin, and Jinsong Su. 2018. Attention regularized sequence-to-sequence learning for e2e nlg challenge. *E2E NLG Challenge System Descriptions*.