Supplementary Material: Long and Diverse Text Generation with Planning-based Hierarchical Variational Model

Zhihong Shao[1], Minlie Huang[1]*, Jiangtao Wen[1], Wenfei Xu[2], Xiaoyan Zhu[1]
[1] Institute for Artificial Intelligence, State Key Lab of Intelligent Technology and Systems
[1] Beijing National Research Center for Information Science and Technology
[1] Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China
[2] Baozun, Shanghai, China
szh19@mails.tsinghua.edu.cn, aihuang@tsinghua.edu.cn
jtwen@tsinghua.edu.cn, xuwenfeilittle@gmail.com
zxy-dcs@tsinghua.edu.cn

A Dataset

A.1 Dataset Statistics

Categories	Attributes	Attribute Values
上衣 (Tops)	衣样式 (Type)	衬衫 (Shirt), 卫衣 (Sweater), 西装 (Suit)
	衣袖型 (Sleeve)	灯笼袖 (Lantern Sleeve), 喇叭袖 (Flare Sleeve), 蝙蝠袖 (Batwing Sleeve)
	衣领型 (Collar)	西装领 (Notch Lapel), Polo领 (Polo-neck Collar), 方领 (Square Collar)
裙 (Dress / Skirt)	裙型 (Type)	连衣裙 (One-piece Dress), 百褶裙 (Pleated Skirt), 伞裙 (Full Skirt)
	裙下摆 (Hem)	荷叶边 (Flounce), 压褶 (Pleated), 垂坠 (Draping)
	裙长 (Length)	长裙 (Long), 半身裙 (Skirt), 超短裙 (Miniskirt)
裤 (Pants)	裤型 (Type)	哈伦裤 (Harem Pants), 铅笔裤 (Pencil Pants), 背带裤 (Overalls)
	裤腰型 (Waist)	高腰 (High-rise), 低腰 (Low-rise), 自然腰 (Regular-rise)
	裤口 (Leg)	开叉 (Split), 毛边 (Cut off), 小脚 (Ankle-tied)

Figure 1: Samples of attributes and attribute values for tops, dress / skirt, and pants.

Figure 1 shows some samples of attributes and attribute values from our dataset for advertising text generation.

A.2 Comparisons with other Benchmark Datasets

Compared with other widely used data-to-text datasets (i.e., WEATHERGOV, WIKIBIO, E2E, WebBLG, and ROTOWIRE), our dataset is more suitable for long and diverse text generation. Table 1 presents the comparisons among the datasets. WEATHERGOV (Liang et al., 2009) defines a task to generate a weather forecast from weather statistics; it has a vocabulary size of less than 400, which indicates the simplicity of its language expressions. WIKIBIO (Lebret et al., 2016) defines a task that requires to generate a very short biography (about 26 words) from personal information. E2E (Novikova et al., 2017) and WebNLG (Gardent et al., 2017) are not confined to one specific domain, but also consist of very short texts (one sentence for most of the time). These datasets are not suitable for long text generation. ROTOWIRE (Wiseman et al., 2017) defines a task that requires to generate a summary from the records of a basketball game; the texts in ROTOWIRE are long enough but lack language expression diversity. We evaluated expression diversity with distinct-4. Before computing distinct-4, we first replaced numbers and named entities with their categories because such words contribute much to distinct-4 but not to expression diversity. For example, *Lakers* was replaced with the tag <*TEAM*> in ROTOWIRE, and *Dior* was replaced with the tag <*BRAND*> in our dataset. We then computed distinct-4 on 3,000 randomly sampled texts from our dataset and on samples with comparable number of words from E2E, WebNLG, WIKIBIO, and ROTOWIRE respectively. Our dataset exhibits much higher diversity with a substantially higher distinct-4 score than other corpora. As the texts in our dataset are long (the av-

*Corresponding author: Minlie Huang.

	E2E	WebNLG	WG	WB	RW	Ours
Avg. Len.	14.3	22.69	28.7	26.1	337.1	110.2
Distinct-4	35.44%	46.84%	N/A	47.04%	48.12%	85.35%
Vocab	65.7K	8K	394	400K	11.3K	54.9K
# Instances	50.6K	25.3K	22.1K	728K	4.9K	119K

Table 1: Statistics of E2E, WebNLG, WEATHERGOV (WG), WIKIBIO (WB), ROTOWIRE (RW) and our dataset. We computed distinct-4 (see section 5.3) on 3,000 randomly sampled advertising texts for our dataset and on samples with comparable number of words for E2E, WebNLG, WB and RW, respectively.

erage length is about 110 words) and diverse, our dataset is suitable to evaluate long and diverse text generation in this paper.

Some other datasets pair structured data with user-generated reviews, such as Amazon reviews (McAuley et al., 2015), Yelp dataset[1], and IMDb dataset (Maas et al., 2011). We did not use such corpora because the contents of user-generated reviews do not mainly come from the data but commonly depend on many other things such as the reviewers' experience and preference.

B Case Study

B.1 Advertising Text Generation

Figure 2 shows generated texts from different models given the same input.

Most baselines fail to cover all the provided data and repeatedly describe some of the input items. For example, the text from Link-S2S ignores the attribute value *three-quarter sleeve* and describes the *round collar* twice. Checklist and CVAE also have similar problems. As Link-S2S and Checklist inject variations only at the conditional output distribution, they suffer from the redundancy problem. Though Pointer-S2S covers all attribute values without redundancy, it introduces logical incoherence (the *round collar* can not *reveal slender arms*) in the first sentence. By contrast, both texts generated by our model cover all the input data without redundancy.

Due to diverse yet reasonable planning, the two texts of our model exhibit different discourse structures. The first text adopts a general-to-specific discourse structure where the statement in the beginning (i.e., *the elegance of the dress*) is supported by the following descriptions of local features. It groups global features (i.e., *color*, *material* and *length*) from the input in the first sentence and realizes each of the remaining sentences with one local feature. The second text adopts a parallel structure which splits global features and arranges some of them in the middle. Despite the difference, the two texts exhibit a global pattern in the data. They both describe the dress from top to bottom (i.e., *collar* $->$ *sleeve* $->$ *shape of the lower part*), which verifies the effectiveness of content organization. Noticeably, the two texts show diverse wording, which exemplifies that our model captures the diversity of expressions.

B.2 Recipe Text Generation

Figure 3 shows the generated examples. Although the three models fail to cover all given ingredients, our model gives the most complete procedure for making a *pumpkin pie* which includes five steps: *1.beat eggs* $->$ *2.blend with some other ingredients* $->$ *3.pour into pie shell* $->$ *4.bake* $->$ *5.cool*. Our model also gives the most specific and precise instructions for step 4 and step 5. By contrast, all baselines miss step 3 or step 5. Checklist produces the general phrase "combine all of the ingredients". CVAE suffers from the redundancy and incoherence problems. Pointer-S2S mentions the most ingredients but misses the most important one "pumpkin". Link-S2S misses "pumpkin" and generates incoherent expressions.

[1] https://www.yelp.com/dataset/challenge

Input

1 . <类型, 裙>
<Category, Dress / Skirt>
2 . <版型, 显瘦>
<Design, Figure Flattering>
3 . <材质, 蕾丝>
<Material, Lace>
4 . <颜色, 黑色>
<Color, Black>
5 . <风格, 简约>
<Style, Minimalist>
6 . <风格, 性感>
<Style, Sexy>
7 . <裙型, A字>
<Shape, A-line>
8 . <裙长, 长裙>
<Length, Long>
9 . <裙袖长, 七分袖>
<Sleeve Length, Three-quarter Sleeve>
10. <裙领型, 圆领>
<Collar, Round>
11. <裙款式, 拼接>
<Element, Stitching>

Methods	Missing Pairs	texts
Checklist	11	这款黑色蕾丝长裙，简约的圆领设计，修饰颈部曲线，性感迷人。 This dress with black lace, the minimalist round collar flatters the curve of your neck which is sexy and attractive. 七分袖设计，修饰手臂曲线，更显纤细修长。 The three-quarter sleeves flatter the curves of your arms, making your arms look slender. A字版型，遮肉显瘦，不挑身材，适合各种身材。 The A-line shape is figure flattering and fit for all kinds of figures. 七分袖设计，修饰手臂曲线，更显纤细修长。 The three-quarter sleeves flatter the curves of your arms, making your arms look slender.
CVAE	4	这是一款十分简约的A字长裙，采用了经典的圆领设计，经典的圆领，贴合颈部肌肤，带来舒适的穿着体验。 This is a very minimalist A-line dress, which adopts a classic design of round collar, the classic round collar fits your neck and brings you comfortable dressing experience. 七分袖的设计，能够很好的修饰手臂的曲线，带来显瘦的穿着效果。 The design of three-quarter sleeves perfects the curves of your arms with the effect of flattering your figure. 蕾丝拼接的设计，带来了浓郁的性感气息。 The design of lace cutting and stitching makes you look very sexy. 七分袖的设计，穿着舒适自在 。 The design of three-quarter sleeves makes you comfortable.
Pointer-S2S		这款黑色的长裙，简约的圆领七分袖设计，露出纤细的手臂。 This black dress, the minimalist designs of the round collar and three-quarter sleeves reveal your slender arms. 拼接蕾丝的设计，若隐若现的露出白皙的肌肤，性感迷人。 The design of lace cutting and stitching shows a looming presence of your white skin, which is sexy and attractive. A字裙的版型，显瘦又显高。 The A-line shape is figure flattering and makes you look taller.
Link-S2S	9, 11	黑色的蕾丝长裙，简约的圆领设计，修饰脸型，更显性感迷人。 The black dress with lace, the minimalist round collar flatters your face shape, which makes you look sexier and more attractive. A字版型，修身显瘦，展现女性优雅身姿。 The A-line shape flatters your figure and shows your graceful posture. 简约的圆领设计，修饰脸型，更显脸小精致。 The minimalist round collar flatters your face shape, which makes your face look more delicate.
PHVM		Groups / Sentences
		3, 4, 8, 11: 黑色的蕾丝拼接长裙，展现出别具一格的气质。 This is a black dress with lace cutting and stitching, which has an aura of unique elegance.
		6, 10: 圆领的设计既能露出女性性感的锁骨，又能起到修饰脸型的作用。 The round neckline not only shows off your sexy collarbones, but also flatters your face shape.
		5, 9: 简约的七分袖，修饰纤细的手臂，散发出时尚气息。 The minimalist three-quarter sleeves betray your slender arms, radiating fashion.
		2, 7: 修身的A字版型，提高了腰线，更有显瘦的效果 。 The A-line shape heightens your waistline visually, which is figure flattaring.
		2: 略微修身的版型，能很好地修饰身材，穿着舒适又显瘦。 Its lightly slimming design flatters your figure, while keeping your body comfortable.
		3, 4, 6, 11: 黑色蕾丝拼接的点缀，性感时尚，亮眼吸睛。 The decoration of black lace cutting and stitching is sexy, fashionable and buzzworthy.
		5, 10: 简约的小圆领，搭配七分袖处理，举手投足间尽显优雅气质。 The minimalist round collar, paired with the three-quarter sleeves, shows your elegant temperament.
		7, 8: A字长裙版型，上身凸显纤细腰肢，让你轻松穿出女神范。 The dress's shape is A-line, which accentuates your slender waist and endows you a beautiful look.

Figure 2: Generated advertising texts from different models. Attribute values are colored in red. Repeated expressions are underlined.

Title:	Drunken apple/pumpkin pie	
Ingredients:	2 eggs 3/4 c brown sugar ; firmly packed 1 9-inch pie shell ; unbaked 1 c cooked pumpkin - mashed and drained 1 c applesauce ; thick & chunky 1 tb flour 1/2 ts salt	1 ts each cinnamon and ginger 1/4 ts nutmeg 1/8 ts each allspice and cloves 1 1/2 c half-and-half or vaporated milk 1 ts vanilla 1 c pecan halves 2 tb rum
Truth	In a bowl, beat together the eggs and sugar until light. Mix in the pumpkin, applesauce, flour, salt, cinnamon, ginger, nutmeg, allspice, cloves, half-and-half and vanilla; Blend thoroughly. Pour into pie shell. Arrange pecan halves over top of filling. And bake 30 to 35 minutes longer , or until filling is firm and a knife inserted 1 " from the edge comes out clean. Cool on a wire rack. At serving time, warm rum in a small container suitable for pouring. Light the rum with a match and pour immediately while flaming over the pie.	
Checklist	Preheat oven to 350 degrees. Grease a 9 " pie plate. In large bowl, combine all of the ingredients, mixing well. Pour into a greased 9x13 " pan. Bake for 1 hour at 350 degrees for 1 hour. Cool. Store in airtight container.	
CVAE	Beat the eggs with the sugar, flour, salt, cinnamon and nutmeg in a bowl. Make a well in the center of the mixture and set aside. Combine the pumpkin, flour, salt, cinnamon, nutmeg, and nutmeg and stir into the eggs. Add the half-and-half, flour, salt, and nutmeg, mixing well. Bake in a 350 f oven for 45 minutes, or until a knife inserted in the center comes out clean. Cool in the pan for 5 minutes, then remove from the pan.	
Pointer-S2S	Beat eggs until light and fluffy. Add brown sugar, applesauce, flour, salt, nutmeg, cloves, allspice, and pecans. Pour into pie shell. Bake at 350 degrees for 45 minutes.	
Link-S2S	Beat eggs until light and fluffy. Add sugar, eggs, and vanilla. Mix well. Add eggs, one at a time, beating well after each addition. Stir in the flour, baking powder, salt, and vanilla. Pour into a greased and floured 9 " x 13 " pan. Bake at 350 degrees for 45 minutes.	
PHVM	In a large mixing bowl , beat together the eggs and brown suger until light and fluffy. Stir in the pumpkin. Blend in the flour, salt and nutmeg or half-and-half. Stir in vanilla and pecans. Pour into prepared pie shell. Bake for 30-35 minutes or until toothpick inserted in the center comes out clean. Cool on racks.	

Figure 3: Generated recipes from different models.

References

Claire Gardent, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. 2017. Creating training corpora for NLG micro-planners. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 179–188.

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, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 1203–1213. The Association for Computational Linguistics.

Percy Liang, Michael I. Jordan, and Dan Klein. 2009. Learning semantic correspondences with less supervision. In *ACL/IJCNLP*, pages 91–99. The Association for Computer Linguistics.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *ACL*, pages 142–150. The Association for Computer Linguistics.

Julian J. McAuley, Christopher Targett, Qinfeng Shi, and Anton van den Hengel. 2015. Image-based recommendations on styles and substitutes. In *SIGIR*, pages 43–52. ACM.

Jekaterina Novikova, Ondrej Dusek, and Verena Rieser. 2017. The E2E dataset: New challenges for end-to-end generation. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue, Saarbrücken, Germany, August 15-17, 2017*, pages 201–206.

Sam Wiseman, Stuart M. Shieber, and Alexander M. Rush. 2017. Challenges in data-to-document generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 2253–2263. Association for Computational Linguistics.