SAPPHIRE: Approaches for Enhanced Concept-to-Text Generation

Carnegie Mellon University

Steven Y. Feng, Jessica Huynh, Chaitanya Narisetty, Eduard Hovy, Varun Gangal
Language Technologies Institute, Carnegie Mellon University

1. Summary

- **Motivation:** Seeking simple and effective improvements for concept-to-text generation
- **Focus:** CommonGen or generative commonsense reasoning task, which involves generating logical sentences from a given set of input concepts
- **SAPPHIRE:** Set Augmentation and Post-hoc PHrase Infilling and Recombination
  - Concept set augmentation based on keywords and attention
  - Phrase recombination for generating more logical and coherent sentences

2. CommonGen: Overview and Baselines

- **Task:** input concept set → output logical sentence. Examples:
  - *{horse, carriage, draw}* → *The carriage is drawn by the horse.*
  - *{listen, talk, sit}* → *The man told the boy to sit down and listen to him talk.*
- **Dataset:** created new dev, test splits (dev_{CG}, test_{CG}) from original dev set (dev_{O}) for our experiments since original test set (test_{O}) is hidden. Training set (train_{CG}) was unaltered

Stats	Train_{CG}	Dev_{O}	Test_{O}	Dev_{CG}	Test_{CG}
# concept sets	32,651	993	1,497	240	360
# sentences	67,389	4,018	7,644	984	1583

- **Baselines:** trained 4 seq2seq Transformer models – BART-base, BART-large, T5-base, T5-large. Performance of our re-implemented models exceeded original reported scores

3. Thorough Baseline Analysis

- **Correlation Study:**

Question	• Does the number of input concepts affect the quality of generated text?
Observations	• Most metrics are positively correlated with concept set size • ROUGE-L, CIDEr, SPICE have statistically insignificant correlations • Coverage is strongly negatively correlated with concept set size
Takeaways	• Increased concept set size results in greater overall performance • Probability of concepts missing from generated text increases with concept set size

- **Qualitative Analysis:** Issues observed in generated baseline texts are listed below
  - Sometimes lack commonsense and/or fluency, i.e. outputs often seem more like phrases than fully coherent sentences
  - Can miss important words, e.g. "*A listening music and dancing in a dark room*"
  - Generally generic and bland, e.g. "*Someone sits and listens to someone talk*"
  - Improper ordering of sentence segments, e.g. "*body of water on a raft*"

4. SAPPHIRE

4.1 Concept Set Augmentation

- Motivated by the correlation study to improve performance and coverage, we augment concept sets with additional words (from 1 to 5 words) as new inputs to the models
- During training, additional keywords are extracted from the human references
- During inference, they are extracted from the baseline model generations

- **Keyword-based Augmentation (*Kw-aug*):**
  - Use KeyBERT to extract keywords from the texts
  - Calculate average semantic similarity of candidate keywords with original concept set
  - Add remaining candidate with highest similarity at each augmentation stage

Original Concept Set	Added Words
{match, stadium, watch}	{soccer, league, fans}
{family, time, spend}	{holidays}
{head, skier, slope}	{cabin}

- **Attention-based Augmentation (*Att-aug*):**
  - Pass texts through BERT and return the attention weights at the last layer
  - Identify words in the text that are most attended upon in aggregate
  - Add remaining candidate with the highest attention at each augmentation stage

Original Concept Set	Added Words
{boat, lake, drive}	{fisherman}
{family, time, spend}	{at, holidays}
{player, match, look}	{tennis, on, during}

4. SAPPHIRE

4.2 Phrase Recombination

- Motivated by the qualitative analysis, we break down sentences into phrases and reconstruct them (plus original concepts) into new sentences with more coherence
- During training, YAKE is used to extract phrases (2,3,5 n-grams) from human references
- During inference, YAKE is used to extract keyphrases from baseline model generations

- **Phrase-to-text (P2T):**
  - Trains the models to become order-agnostic by piecing the phrases back together
  - Input: random permutation of keyphrases + concepts → output: human references
  - During inference, a single random permutation of keyphrases + concepts as input

- **Mask Infilling (MI):**
  - Interpolates text between test-time input set elements with no training required
  - Given an input set $\{c_1, c_2\}$, we feed "$< mask > c1 < mask > c2 < mask >$" and "$< mask > c2 < mask > c1 < mask >$" to an MI model (here, we use BART)
  - Difficulty: determining the best input set permutation to produce good output text
  - Proposal: use perplexity (PPL) from GPT-2 to select the *best* permutations

Original Text	Extracted Keyphrases	New Input Concept Set
A dog wags his tail at the boy.	dog wags his tail	{dog wags his tail}
hanging a painting on a wall at home	hanging a painting	{hanging a painting, wall}
A herd of many sheep crowded together in a stable waiting to be dipped for ticks and other pests	herd of many sheep crowded	{herd of many sheep crowded, dip, waiting}

5. Experiments and Results

- Epochs with best ROUGE-2 score on the dev split are chosen for beam-search decoding
- Human evaluation of fluency and commonsense on 1-5 scales for human references, baseline generations, and SAPPHIRE model outputs for BART-large and T5-base

- Automatic evaluation results on test_{CG}

	BART-base					BART-large				
Metrics\Methods	Baseline	Kw-aug	Att-aug	P2T	BART-base-MI	Baseline	Kw-aug	Att-aug	P2T	BART-large-MI
ROUGE-1	43.96±0.03	**45.01**±0.00	44.99±0.10	44.87±0.42	44.83	45.67±0.25	46.71±0.05	**46.78**±0.14	46.26±0.29	41.69
ROUGE-2	17.31±0.02	**18.33**±0.06	18.18±0.04	18.04±0.13	17.44	18.77±0.04	19.64±0.05	**19.92**±0.19	19.37±0.17	15.40
ROUGE-L	36.65±0.00	37.28±0.24	**37.76**±0.12	37.28±0.11	34.47	37.83±0.29	38.38±0.01	**38.53**±0.03	38.22±0.16	33.32
BLEU-1	**73.20**±0.28	73.00±0.85	73.00±0.14	73.15±1.06	69.90	74.45±0.21	76.20±0.99	76.55±0.92	**77.10**±0.85	63.90
BLEU-2	54.50±0.14	55.35±0.49	**55.70**±0.28	55.65±0.35	49.00	56.25±0.78	58.60±0.14	**59.60**±0.00	58.95±0.64	42.40
BLEU-3	40.40±0.14	41.35±0.21	41.40±0.28	**41.85**±0.35	34.70	42.15±0.49	44.00±0.00	**45.20**±0.42	44.70±0.14	29.20
BLEU-4	30.10±0.14	31.10±0.14	30.95±0.07	**31.75**±0.35	24.70	32.10±0.42	33.40±0.28	**34.50**±0.42	34.25±0.21	20.50
METEOR	30.35±0.35	30.50±0.28	30.70±0.14	**31.05**±0.49	29.70	31.70±0.14	32.60±0.57	32.65±0.49	**33.00**±0.14	28.30
CIDEr	15.56±0.10	**16.18**±0.12	15.68±0.00	16.14±0.33	14.43	16.42±0.09	17.37±0.08	17.49±0.49	**17.50**±0.02	12.32
SPICE	30.05±0.07	30.45±0.07	30.65±0.35	**30.95**±0.21	28.40	31.85±0.21	33.15±0.49	33.30±0.28	**33.60**±0.00	26.10
BERTScore	59.19±0.32	59.32±0.25	**59.72**±0.03	59.54±0.05	53.73	59.95±0.29	60.83±0.29	60.87±0.45	**61.30**±0.66	48.56
Coverage	90.43±0.17	91.44±0.95	91.23±0.21	91.47±2.93	**96.23**	94.49±0.53	96.74±1.20	96.02±1.17	**97.02**±0.15	95.33

	T5-base					T5-large				
Metrics\Methods	Baseline	Kw-aug	Att-aug	P2T	BART-base-MI	Baseline	Kw-aug	Att-aug	P2T	BART-large-MI
ROUGE-1	44.63±0.13	46.42±0.01	**46.75**±0.11	45.73±0.27	44.92	46.26±0.17	**47.47**±0.16	47.40±0.12	46.72±0.26	42.78
ROUGE-2	18.40±0.14	**19.36**±0.24	19.20±0.17	18.51±0.11	17.98	19.62±0.17	20.02±0.07	**20.19**±0.01	19.76±0.22	16.61
ROUGE-L	37.60±0.16	**38.68**±0.08	38.51±0.21	38.07±0.10	34.88	39.21±0.22	39.84±0.12	**39.97**±0.06	39.19±0.09	34.52
BLEU-1	73.60±0.85	**76.25**±0.35	76.00±0.28	75.65±1.20	70.20	77.45±0.21	78.70±0.28	**78.95**±0.07	77.90±0.57	66.80
BLEU-2	57.00±0.71	**59.55**±0.64	58.75±0.35	58.15±0.64	50.50	60.75±0.21	62.10±0.14	**62.35**±0.07	61.00±0.42	45.90
BLEU-3	42.75±0.49	**45.10**±0.85	44.00±0.28	43.45±0.07	36.20	46.60±0.14	47.65±0.21	**47.95**±0.21	46.75±0.49	32.70
BLEU-4	32.70±0.42	**34.45**±0.92	33.30±0.28	33.10±0.28	26.10	36.30±0.00	36.80±0.28	**37.35**±0.49	36.10±0.42	23.90
METEOR	31.05±0.49	31.85±0.07	31.90±0.14	**32.05**±0.35	30.20	33.30±0.14	33.55±0.07	**33.70**±0.00	33.35±0.21	29.10
CIDEr	16.26±0.25	**17.37**±0.04	17.04±0.21	16.84±0.11	14.83	17.90±0.15	18.40±0.18	**18.43**±0.10	17.89±0.08	13.34
SPICE	31.95±0.07	32.75±0.21	32.85±0.21	**33.20**±0.14	29.70	34.25±0.07	**34.50**±0.28	33.70±0.14	34.00±0.28	28.00
BERTScore	61.40±0.34	**61.88**±0.06	61.28±0.10	61.46±0.01	55.04	62.65±0.07	**62.91**±0.15	62.78±0.21	62.46±0.11	50.57
Coverage	90.96±1.77	94.92±0.45	96.00±0.03	94.78±0.83	**96.03**	94.23±0.21	95.92±0.02	**96.08**±0.09	95.44±0.58	96.03

- Automatic evaluation results on hidden test_{O} (evaluated by the CommonGen authors)

Models\Metrics	ROUGE-2/L		**BLEU-3/4**		METEOR	**CIDEr**	**SPICE**	Coverage
T5-base (reported baseline)	14.63	34.56	28.76	18.54	23.94	9.40	19.87	76.67
BART-large (reported baseline)	22.02	41.78	39.52	29.01	31.83	13.98	28.00	97.35
T5-large (reported baseline)	21.74	42.75	43.01	31.96	31.12	15.13	28.86	95.29
EKI-BART (Fan et al., 2020)	-	-	-	35.945	-	16.999	29.583	-
KG-BART (Liu et al., 2021)	-	-	-	33.867	-	16.927	29.634	-
RE-T5 (Wang et al., 2021)	-	-	-	**40.863**	-	**17.663**	**31.079**	-
BART-base-P2T	20.83	42.91	40.74	29.918	30.61	14.670	26.960	92.84
T5-base-P2T	22.38	44.59	44.97	33.577	31.95	16.152	29.104	95.55
BART-large-KW	22.25	43.38	43.87	32.956	32.26	16.065	28.335	96.16
BART-large-Att	22.22	43.80	44.61	33.405	32.03	16.036	28.452	96.43
BART-large-P2T	22.65	43.84	44.78	33.961	32.18	16.174	28.462	96.20
T5-large-KW	23.79	45.73	48.06	37.023	32.85	16.987	29.659	95.32
T5-large-Att	23.94	45.87	47.99	36.947	32.79	16.943	29.607	95.43
T5-large-P2T	23.89	45.77	48.08	37.119	32.94	16.901	29.751	94.82

- Human evaluation results on test_{CG}

Model	Method	**Fluency**	**Commonsense**
BART-large	Baseline	3.92	4.06
	Kw-aug	4.13	3.92
	Att-aug	4.10	4.06
	P2T	**4.17**	**4.13**
T5-base	Baseline	4.02	3.83
	Kw-aug	4.04	4.04
	Att-aug	**4.13**	3.98
	P2T	4.02	**4.08**
Human		4.14	4.32

- Qualitative example with model outputs

Concept Set	{sit, chair, toy, hand}
Baseline	hands sitting on a chair
Kw-aug	A boy sits on a chair with a toy in his hand.
Att-aug	A child sits on a chair with a toy in his hand.
P2T	Hands sitting on a chair with toys.

6. Conclusion and Future Work

- Proposed several improvements called SAPPHIRE for concept-to-text generation
- Demonstrated its effectiveness thoroughly on the CommonGen task with BART and T5
- Possible to explore various combinations of proposed SAPPHIRE methods
- Also possible to try improving the performance of the mask infilling approach
- Can study SAPPHIRE on other data-to-text tasks like WebNLG, for dialog agents, etc.

Email: syfeng@cs.cmu.edu