POINTER: Constrained Progressive Text Generation via Insertion-based Generative Pre-training

Zhang et al., EMNLP 2020

Outline

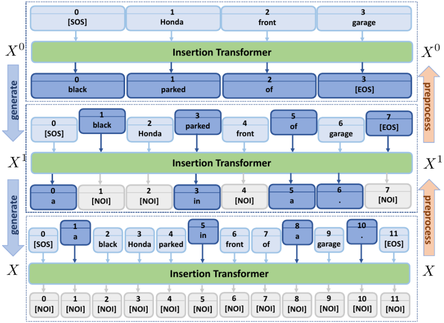
- This paper introduces a **hard-constrained**, **progressive insertion** text generation method, learned via insertion-based generative pre-training
  - Given a set of key words, the model will insert words in between these key words, establish a new set of key words, the model keep repeat the this process until a complete sentence is formed (satisfying certain condition).
  - Desired property: most important (informative) words (such as nouns, verbs, rare words) are inserted first then lesser important words (such as stop words and high-frequent words)
- The model is pre-trained on Wikipedia dataset, then fine-tuned on downstream datasets for usage, such as WMT News and Yelp.

Stage	Generated text sequence
0 (X^0)	sources sees structure perfectly
1 (X^1)	sources **company** sees **change** structure perfectly **legal**
2 (X^2)	sources **suggested** company sees **reason** change **tax** structure **which** perfectly legal **.**
3 (X^3)	**my** sources **have** suggested **the** company sees **no** reason **to** change **its** tax structure **,** which **are** perfectly legal .
4 (X^4)	my sources have suggested the company sees no reason to change its tax structure , which are perfectly legal .

Taxonomy of Text Generation

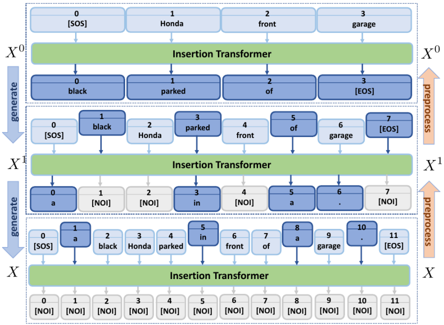
- Soft- vs Hard-constrained text generation models:
  - Soft-constrained models:
    - Approach: conditional text generation based on given set of key words (with other conditioning information) $P(X) = \prod_t^T P(x_t|X_{\ t}, Keys)$
    - These models donot generate exact key words, they can just generate similar words to the key words
  - Hard-constrained models:
    - Approach: often construct a lexical-constrained grid beam search decoding algorithm to incorporate the set of key words
- Autoregressive vs Progressive text generation models:
  - Autoregressive models
    - Generating each word based on previous words, e.g. GPT models with causal language modeling
  - Progressive (or non-autoregressive) models:
    - Generating words simultaneously based on other words, e.g. BERT models with masked language modeling
- This paper introduces a **hard-constrained**, **progressive insertion** text generation method

POINTER Model: Generation (Inference) Procedure



- Multi-stage generation:
  - At each stage X^k, given a set of words, the model will apply seq2seq prediction for this set of words
  - Predicted words (except special token $[NOI]$) will be inserted between previous words.
  - We repeat this process until no additional word is generated, in other words, once in a stage which all slots predict all $[NOI]$

POINTER Model: Generation (Inference) Procedure



- Desired property: the model will insert more important (informative) words (such as nouns, verbs, rare words) in earlier stages, then insert lesser important words (such as stop words and high-frequent words) in later stages
- An issue: the model may new repeating words at each stage e.g. from a current word "and", the model generates "clean and clean"
  - Inner-layer beam search (ILBS), which aims to select most satisfying words from the C best predicting candidates.

POINTER Model: Training and Data Preparation

- Training: based on the inference procedure, the model will be trained via seq2seq objective (most likely)
- Data preparation: reversing the generation process
  - We construct pairs of text sequences at adjacent stages i.e., (X^{k-1}, X^k),
  - Each training instance full sentence X is broken into a consecutive series of pairs $(X^0, X^1), \ldots, (X^{K-1}, X^K)$,
  - According a dropping criteria, we reversely drop non-consecutive words from the full sentence at each stage, $X^K \rightarrow X^{K-1}$, until only $4-7$ key words remaining
- Dropping criteria and formula, solved by a DP algorithm
  - Reversing the generating process, we drop lesser important (informative) words earlier, drop more important later
  - We do not drop consecutive words at any stage

$$max \sum_{t}^{T} \phi_t(\alpha_{\texttt{max}} - \alpha_t)$$

$$\texttt{s.t.} \quad \phi_t \phi_{t+1} \neq 1, \forall t$$

$$\texttt{where} \quad \alpha_{\texttt{max}} = max_t\{\alpha_t\}, \quad \phi_t \in \{0, 1\}$$

Experimental Results

News dataset Method	NIST N-2	N-4	BLEU B-2	B-4	METEOR	Entropy E-4	Dist D-1	D-2	PPL.	Avg. Len.
CGMH	1.60	1.61	7.09%	1.61%	12.55%	9.32	**16.60%**	**70.55%**	189.1	14.29
NMSTG	2.70	2.70	10.67%	1.58%	13.56%	10.10	11.09%	65.96%	171.0	27.85
Greedy (base)	2.90	2.80	12.13%	1.63%	15.66%	**10.41**	5.89%	39.42%	97.1	47.40
Greedy (+Wiki,base)	3.04	3.06	13.01%	2.51%	**16.38%**	10.22	11.10%	57.78%	56.7	31.32
ILBS (+Wiki,base)	3.20	3.22	14.00%	2.99%	15.71%	9.86	13.17%	61.22%	66.4	22.59
Greedy (+Wiki, large)	**3.28**	**3.30**	**14.04%**	**3.04%**	15.90%	10.09	12.23%	60.86%	**54.7**	27.99
Human oracle	-	-	-	-	-	10.05	11.80%	62.44%	47.4	27.85

Yelp dataset Method	NIST N-2	N-4	BLEU B-2	B-4	METEOR	Entropy E-4	Dist D-1	D-2	PPL.	Avg. Len.
CGMH	0.50	0.51	4.53%	1.45%	11.87%	9.48	**12.18%**	**57.10%**	207.2	16.70
NMSTG	1.11	1.12	10.06%	1.92%	13.88%	10.09	8.39%	50.80%	326.4	27.92
Greedy (base)	2.15	2.15	11.48%	2.16%	**17.12%**	**11.00**	4.19%	31.42%	99.5	87.30
Greedy (+Wiki,base)	3.27	3.30	15.63%	3.32%	16.14%	10.64	7.51%	46.12%	71.9	48.22
ILBS (+Wiki,base)	3.34	3.38	16.68%	3.65%	15.57%	10.44	9.43%	50.66%	61.0	35.18
Greedy (+Wiki, large)	**3.49**	**3.53**	**16.78%**	**3.79%**	16.69%	10.56	6.94%	41.2%	**55.5**	48.05
Human oracle	-	-	-	-	-	10.70	10.67%	52.57%	55.4	50.36

Generated Examples

Keywords	**estate pay stay policy**
CGMH	an economic **estate** developer that could **pay** for it is that a **stay policy** .
NMSTG	as **estate** owners , they cannot **pay** for households for hundreds of middle - income property , buyers **stay** in retail **policy** .
POINTER (Greedy, base)	if you buy new buildings from real **estate** company, you may have to **pay** down a mortgage and **stay** with the **policy** for financial reasons .
POINTER (ILBS, base)	but no matter what foreign buyers do , real **estate** agents will have to **pay** a small fee to **stay** consistent with the **policy** .
POINTER (Greedy, Large)	but it would also be required for **estate** agents , who must **pay** a larger amount of cash but **stay** with the same **policy** for all other assets .

Table 3: Generated examples from the News dataset.

Keywords	**joint great food great drinks greater staff**
CGMH	very cool **joint** with **great food** , **great drinks** and even **greater staff** . ! .
NMSTG	awesome **joint** . **great** service. **great food** great **drinks**. good to **greater** and great **staff**!
POINTER (Greedy, base)	my favorite local **joint** around old town. **great** atmosphere, amazing **food**, delicious and delicious coffee, **great** wine selection and delicious cold **drinks**, oh and maybe even a **greater** patio space and energetic front desk **staff**.
POINTER (ILBS, base)	the best breakfast **joint** in charlotte . **great** service and amazing **food** . they have **great** selection of **drinks** that suits the **greater** aesthetic of the **staff** .
POINTER (Greedy, Large)	this is the new modern breakfast **joint** to be found around the area . **great** atmosphere , central location and excellent **food** . nice variety of selections . **great** selection of local craft beers , good **drinks** . quite cheap unless you ask for **greater** price . very friendly patio and fun **staff** . love it !

Table 4: Generated examples from the Yelp dataset.

Thank you !