A Language Model based Evaluator for Sentence Compression

Yang Zhao[1], Zhiyuan Luo[1], Akiko Aizawa[2]
[1]The University of Tokyo, Japan, [2]National Institute of Informatics, Japan

Sentence Compression

Task: delete words in the **original sentence** to form short **compression** that remains readable and preserves its underlying meaning.

Original Sentence: *ABCDEFGHIJ.*
Compression: *AB~~CDE~~FG~~HI~~J.*

Example Sentence: *A man suffered a* serious *head injury after a* morning car *crash* today.

Example Compression: *A man suffered a head injury after a crash.* _

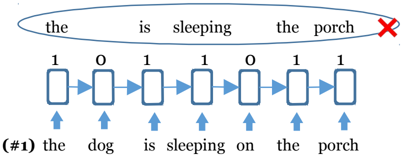
Research Question – How to optimize grammaticality of generated compression ?

(a) Rule-based approaches is too general, sometimes yielding wrong compression.

e.g. Rule - delete prepositional phrase in syntactic tree.

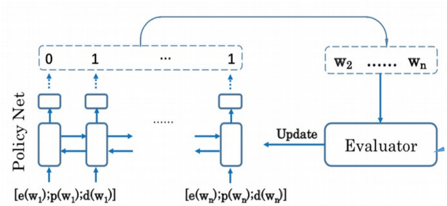
(#1) The dog is sleeping ~~on the porch~~. ✓
(#2) ~~At the heart of the problem~~ is a complex physical model proposed by professor Paul. ✗

(b) Max-likelihood training objective might not guarantee the readability of the compression.

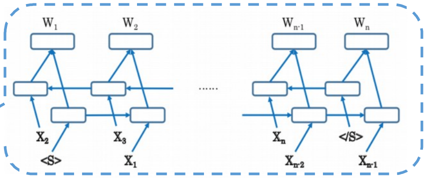


Proposed Approach - A Language Model based Evaluator

Reinforcement Learning framework



Syntax-based Language Model (**SLM**)



$$R_{SLM}(\hat{Y}) = e^{\left(\frac{1}{|\hat{Y}|}\sum_{t=1}^{|\hat{Y}|} logP_{LM}(y_t|y_{0:t-1})\right)}$$

$e(w_i)$: word embedding; $p(w_i)$: part-of-speech embedding; $d(w_i)$: dependency label embedding;
Total Reward $= R_{SLM} + R_{CR}$ where R_{CR} is the compression rate reward. The total reward is used to reinforce the policy network in search of the best compression.

Evaluation and Results

Gigaword Dataset	Annotator 1		Annotator 2		
	F_1	**RASP-F_1**	F_1	**RASP-F_1**	**CR**
#1 Seq2seq with attention	54.9	60.3	58.6	64.6	0.53
#2 Dependency tree+ILP	58.0	65.1	61.0	70.9	0.55
#3 LSTMs+pseudo label	60.3	64.1	64.1	69.2	0.51
#4 Evaluator-LM	64.5	67.3	66.9	72.2	0.50
#5 Evaluator-SLM	**65.0**	**69.6**	**68.2**	**73.9**	0.51

Google Dataset	F_1	**RASP-F_1**	**CR**
&1 Seq2seq with attention	71.7	63.8	0.34
&2 LSTM (Filippova, 2015)	82.0	-	0.38
&3 LSTMs (our implement)	84.8	81.9	0.40
&4 Evaluator-LM	85.0	82.0	0.41
&5 Evaluator-SLM	85.1	82.3	0.39

Table 1: Automatic evaluations (F_1 & RASP-F_1). #1, #2, and #3 are comparison methods while #4 and #5 are proposed methods.

Gigaword	*Readability*	*Informativeness*
$1 LSTMs	3.56	3.10
$2 SLM	**4.16**†	3.16

Table 2: Human evaluation for Gigaword dataset. 5-point scale is used.

(1) The proposed methods yield better compression upon automatic evaluation (Table 1; F_1 $ RASP-F_1).

(2) Compared with Evaluator-LM, Evaluator-SLM brings further improvement, suggesting syntax feature may enable model to learn more unseen but reasonable word collocations. (e.g. noun is usually followed by verb rather than adjective).

(3) Readability of compression is improved by the proposed evaluator upon the human evaluation (Table 2), showing that evaluator-SLM could be used as a post hoc grammar checker.

Contact: Yang Zhao, zhao@is.s.u-tokyo.ac.jp
The 56th Annual Meeting of the Association for Computational Linguistics, Melbourne, Australia. July 15 – July 20, 2018.