Generic Mechanism for Reducing Repetitions in Encoder-Decoder Models

Ying Zhang[1], Hidetaka Kamigaito[1], Tatsuya Aoki[1],

Hiroya Takamura[2] and Manabu Okumura[1]

[1] Tokyo Institute of Technology

[2] National Institute of Advanced Industrial Science and Technology (AIST)

Content

- Motivation
- Repetition Reduction Module (RRM)
- Experiment
- Results and discussion
- Conclusion

Motivation

- Sequence-to-sequence (seq2seq) models are a dominant paradigm in various natural language generation tasks, such as
    - machine translation
    - text summarization
    - response generation
    - …

Motivation - sample of German-to-English translation task

- As Mi et al. (2016) reported, basic seq2seq models (Bahdanau et al., 2015; Luong et al., 2015) sometimes suffer from a repetition problem.

Source: die einzige wahre wahl war wer, **nicht wann** , und nicht was sie danach taten .

attention

Prediction: the only real choice was who , not when , not when , and not what they did after that .

Target: the only real choice was who , not when , and not what you did after .

Haitao Mi, Baskaran Sankaran, ZhiguoWang, and Abe Ittycheriah. 2016. Coverage embedding models for neural machine translation . In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* , pages 955–960, Austin, Texas. Association for Computational Linguistics

Dzmitry Bahdanau, Kyung Hyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015* .

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, Lisbon, Portugal. Association for Computational Linguistics.

Motivation - sample of response generation task

hi , want to be my friend ?

Gold: hello , sure why not ! how old are you ? i just turned 34

Prediction: sure, **i'm a real estate agent**.

whoa . i am only 12 . what is it like to be old ?

Gold: my 2 kiddos are in kindergarten , they want to graduate at mississippi university like me

Prediction: i'm 35, so i'm not sure. i'm a real estate agent

that is cool i guess . are they twins

Gold: no . only 12 ? so no job yet . i have done real estate 12 years now

Prediction: yes, **they are twins**, they are twins

Motivation

- Previous research for tackling the repetition problem.
    - Constrain the coverage of the attention distribution (Tu et al., 2016; Mi et al., 2016). This method is not suitable for seq2seq models with multiple attentions, such as Transformer (Vaswani et al., 2017).
    - Constrain the upper-bound word frequency or sentence length for the output sentence (Suzuki and Nagata, 2017; Kiyono et al., 2018). It is difficult to apply this method to some tasks other than text summarization.

Zhaopeng Tu, Zhengdong Lu, Yang Liu, Xiaohua Liu, and Hang Li. 2016. Modeling coverage for neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers),* pages 76–85, Berlin, Germany. Association for Computational Linguistics.

Jun Suzuki and Masaaki Nagata. 2017. Cutting-off redundant repeating generations for neural abstractive summarization . In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers* , pages 291–297, Valencia, Spain. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NIPS.*

Shun Kiyono, Sho Takase, Jun Suzuki, Naoaki Okazaki, Kentaro Inui, and Masaaki Nagata. 2018. Reducing odd generation from neural headline generation. In *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation*, Hong Kong. Association for Computational Linguistics.

Content

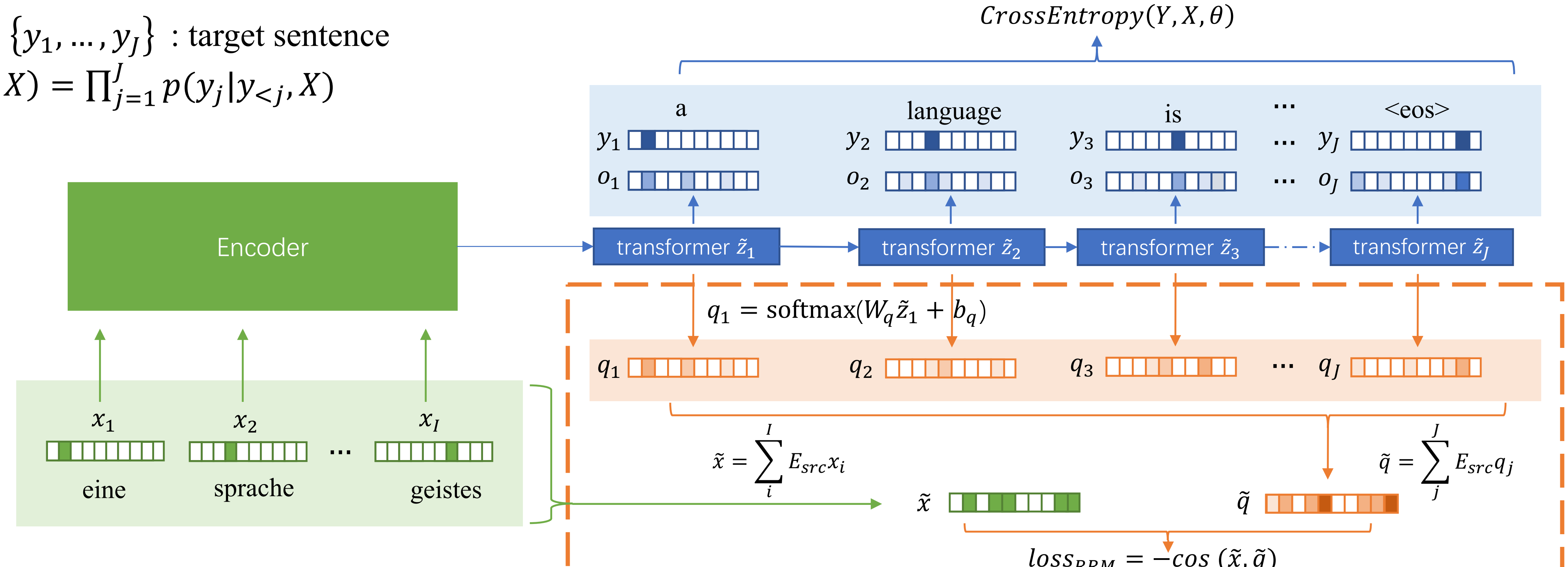
- Motivation
- Repetition Reduction Module (RRM)
- Experiment
- Results and discussion
- Conclusion

Repetition Reduction Module (RRM) - Overview

$X = \{x_1, \dots, x_I\}$: source sentence

$Y = \{y_1, \dots, y_J\}$: target sentence

$p(Y|X) = \prod_{j=1}^{J} p(y_j | y_{<j}, X)$



$p(Y, \tilde{x}|X) = p(\tilde{x}|Y, X)\, p(Y|X)$

$G_t = \sum_{(X,Y) \in D} \{-\log p(Y|X) - \alpha(cos(\tilde{x}, \tilde{q}))\}$

$\tilde{z}_j$: final hidden state

$E_{src} \in R^{H \times |V_s|}$: word embedding matrix

V_s : source vocabulary

H : embedding size

Content

- Motivation
- Repetition Reduction Module (RRM)
- Experiment
- Results and discussion
- Conclusion

Experiment – dataset and baseline

- IWSLT 2014 German-to-English translation task
  - Training:160k
  - Validation:7k
  - Test:7k
    - *Short* (source length $\leq$ 25) : 4927 pairs
    - *Medium* (25< source length $\leq$ 50) : 1524 pairs
    - *Long* (50 < source length) : 299 pairs
  - Baseline: Joint source-target model of Fonollosa et al. (2019)

- PERSONA-CHAT response generation task

  Official dataset of The Conversational Intelligence Challenge 2

  - Training:164k
  - Validation:15k
  - Test:15k
  - Baseline: Transfertransfo model of Wolf et al. (2019)
  - Format of the input sequence: the concatenation of the persona information, up to two turns of history utterances, and the query (the utterance).

Jos´e AR Fonollosa, Noe Casas, and Marta R Costa-juss`a. 2019. Joint source-target self attention with locality constraints. *arXiv preprint arXiv:1905.06596* .

Thomas Wolf, Victor Sanh, Julien Chaumond, and Clement Delangue. 2019. Transfertransfo: A transfer learning approach for neural network based conversational agents. *CoRR*, abs/1901.08149.

Experiment – evaluation metric

- IWSLT 2014 German-to-English translation task
  - tokenized BLEU
  - Meteor
  - Repeat (Kiyono et al., 2018)

Reference : *I like apple and I also like cat.*

Prediction : *I like and also like like dog fish fish.*

	Prediction	Reference	Repeat
I	1	2	0
like	3	2	1
dog	1	0	0
fish	2	0	2
…	…	…	…
SUM	-	-	3

Experiment – evaluation metric

- PERSONA-CHAT response generation task

  - F_1 score

  - Perplexity

  - Repeat was computed by subtracting 1 from the frequency of tokens that occur more than once in the generated sequence.

    Sentence-level: we calculated Repeat only with each generated response.

    Dialog-level: we calculated Repeat with the concatenation of a sequence of the generated responses in a dialog.

Experiment – hyperparameter settings

- IWSLT 2014 German-to-English translation task

We followed the experimental settings of Fonollosa et al. (2019)

Vocab size: 31K
Learning rate: 0.0001
Training steps: 85k
Batch size: 4k mini-batch
Scaling factor α : {1, 0.3, 0.2, 0.05, 0.01}

- PERSONA-CHAT response generation task

We followed the experimental settings of Wolf et al. (2019)

Vocab size: 40K
Learning rate: 6.25×10^{-5}
Training steps: 2 epochs
Batch size: 32 sequences
Scaling factor α : {1, 0.3, 0.2, 0.05, 0.01}

Content

- Motivation
- Repetition Reduction Module (RRM)
- Experiment
- Results and discussion
- Conclusion

Results and discussion - German-to-English results

Model	Repeat	BLEU	Meteor
Fonollosa et al. (2019)*	-	35.70	-
Fonollosa et al. (2019)	1.244	35.61	35.76
+RRM	**1.229**	**35.71**	**35.77**

Table 1: Experimental results on the IWSLT 2014 De-En test dataset.
* indicates the reported score by Fonollosa et al. (2019).

Data	Model	Repeat	BLEU	Meteor
Short	Fonollosa et al. (2019)	**0.552**	37.41	**36.83**
	+RRM	0.554	**37.47**	36.81
Medium	Fonollosa et al. (2019)	2.484	34.28	34.91
	+RRM	**2.467**	**34.38**	**35.00**
Long	Fonollosa et al. (2019)	6.371	33.11	**34.12**
	+RRM	**6.036**	**33.36**	33.99

Table 2: Experimental results on the IWSLT 2014 De-En test dataset at different lengths.

Results and discussion - response generation results

Model	**Repeat (Sentence-Level)**					**Perplexity**	**F_1**
	1-gram	**2-gram**	**3-gram**	**4-gram**	**5-gram**		
Wolf et al. (2019)*	-	-	-	-	-	16.28	19.50
Wolf et al. (2019)	0.755	0.244	0.107	0.056	0.025	**16.31**	18.22
+RRM	**0.699**	**0.210**	**0.090**	**0.045**	**0.018**	16.33	**18.36**

Table 3: Experimental results on the PERSONA-CHAT test dataset. * indicates the reported score by Wolf et al. (2019).

Model	**Repeat (Dialog-Level)**				
	1-gram	2-gram	3-gram	4-gram	5-gram
Wolf et al. (2019)	28.423	14.319	7.786	4.822	2.800
+RRM	**27.952**	**13.982**	**7.605**	**4.743**	**2.791**

Table 4: Experimental results on the PERSONA-CHAT test dataset.

Results and discussion - extensive experiment for response generation task

- First, we investigate whether decoding methods might influence the performance of RRM by comparing beam search with greedy decoding.

- Second, we investigate whether using history utterances in the input sequence for prediction might influence the performance of RRM.

- Third, we investigated the effectiveness of using persona information and the history utterances during training for RRM.
    - *Full*: we use the concatenation of the persona information, the history utterances and the query as a source sentence during training,
    - *Part*: we use only the query as a source sentence during training.
    - *Divide*: we divide the *Full* source sequence into three parts to compute

      $\tilde{x}_p$, W_{q_p} , $\tilde{q}_p$ for the persona information,

      $\tilde{x}_h$, W_{q_h} , $\tilde{q}_h$ for the history utterances ,

      $\tilde{x}_l$, W_{q_l} , $\tilde{q}_l$ for the query

Then, the averaged cosine similarity was calculated between each divided $\tilde{x}$ and $\tilde{q}$.

Results and discussion - extensive experiment for response generation task

Decode	**Model**	**Repeat (Sentence-Level)**					**Perplexity**	**F_1**
		1-gram	2-gram	3-gram	4-gram	5-gram		
Beam	Wolf et al. (2019)*	-	-	-	-	-	16.28	19.50
	Wolf et al. (2019)	0.755	0.244	0.107	0.056	0.025	**16.31**	18.22
	+RRM (α= 0.3, *full*)	**0.699**	**0.210**	**0.090**	0.045	0.018	16.33	**18.36**
	+RRM (α=1, *divide*)	0.746	0.248	0.114	0.063	0.028	16.34	18.20
	+RRM (α=0.2, *part*)	0.703	0.212	**0.090**	**0.043**	**0.017**	16.40	18.27
Beam	Wolf et al. (2019) w/o history	0.902	0.336	0.135	0.067	0.026	**17.96**	**17.30**
	+RRM (α=0.05, *full*)	0.842	0.275	0.100	**0.043**	**0.014**	18.04	17.14
	+RRM (α=1, *divide*)	0.905	0.338	0.146	0.080	0.034	**17.96**	17.16
	+RRM (α=0.2, *part*)	**0.836**	**0.266**	**0.096**	**0.043**	0.015	18.00	17.17
Greedy	Wolf et al. (2019)	1.275	0.477	0.187	0.089	0.037	-	18.02
	+RRM (α=0.3, *full*)	**1.247**	**0.454**	**0.178**	**0.083**	0.034	-	**18.09**
	+RRM (α=1, *divide*)	1.255	0.473	0.199	0.099	0.042	-	17.87
	+RRM (α=0.2, *part*)	1.265	0.469	0.188	0.085	**0.033**	-	18.08

Table 5: The results of the extensive experiments on the PERSONA-CHAT test dataset.

Results and discussion - extensive experiment for response generation task

Decode	Model	Repeat (Dialog-Level)				
		1-gram	2-gram	3-gram	4-gram	5-gram
Beam	Wolf et al. (2019)	28.423	14.319	7.786	4.822	2.800
	+RRM (α= 0.3, *full*)	**27.952**	**13.982**	**7.605**	**4.743**	2.791
	+RRM (α=1, *divide*)	28.034	14.275	7.894	4.955	2.931
	+RRM (α=0.2, *part*)	27.956	14.066	7.663	4.762	2.773
Beam	Wolf et al. (2019) w/o history	33.058	19.399	11.940	7.960	5.180
	+RRM (α=0.05, *full*)	**32.306**	**18.650**	**11.265**	**7.330**	**4.671**
	+RRM (α=1, *divide*)	33.340	19.814	12.347	8.331	5.465
	+RRM (α=0.2, *part*)	32.696	18.934	11.559	7.698	5.040
Greedy	Wolf et al. (2019)	32.960	17.208	8.852	5.022	2.805
	+RRM (α=0.3, *full*)	**32.559**	**16.741**	**8.532**	4.852	2.706
	+RRM (α=1, *divide*)	32.692	17.115	8.872	5.110	2.867
	+RRM (α=0.2, *part*)	32.678	16.919	8.599	**4.822**	**2.689**

Table 7: The results of the extensive experiments on the PERSONA-CHAT test dataset.

Content

- Motivation
- Repetition Reduction Module (RRM)
- Experiment
- Results and discussion
- Conclusion

Conclusion

- We proposed a novel mechanism Repetition Reduction Module to suppress repetitions.
- It is potential to apply our proposal to other Seq2seq models.
- Experimental results on the IWSLT 2014 German-to-English translation task and the PERSONA-CHAT response generation task demonstrated the effectiveness of our proposal.
- The results of the extensive experiments for the response generation task showed RRM has the ability to handle a concatenated input sequence.

THANK YOU!