Best Paper of ACL2019

Bridging the Gap between Training and Inference for Neural Machine Translation

Wen Zhang[1,2] **Yang Feng**[1,2*] **Fandong Meng**[3] **Di You**[4] **Qun Liu**[5]

[1]Key Laboratory of Intelligent Information Processing
Institute of Computing Technology, Chinese Academy of Sciences (ICT/CAS)
[2]University of Chinese Academy of Sciences, Beijing, China
{zhangwen, fengyang}@ict.ac.cn
[3]Pattern Recognition Center, WeChat AI, Tencent Inc, China
fandongmeng@tencent.com
[4]Worcester Polytechnic Institute, Worcester, MA, USA
dyou@wpi.edu
[5]Huawei Noah's Ark Lab, Hong Kong, China
qun.liu@huawei.com

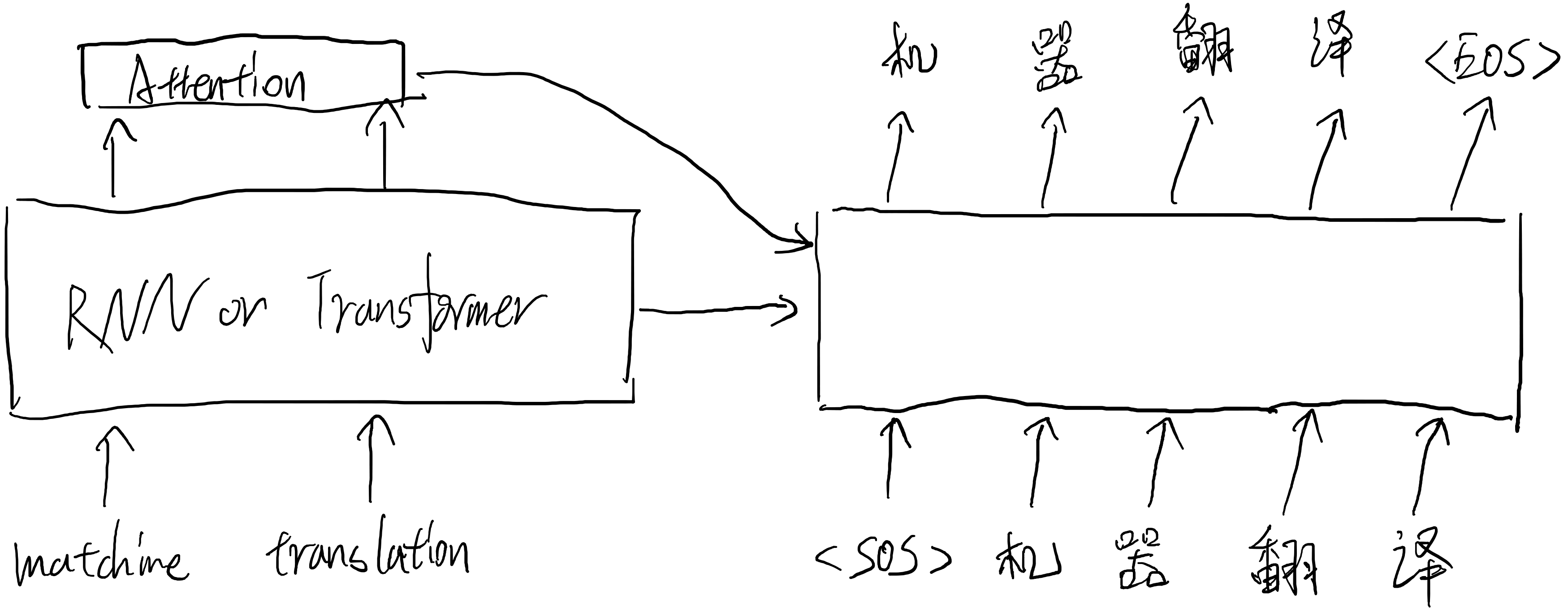
What they do

- Find the problem in decoding → Teacher Forcing

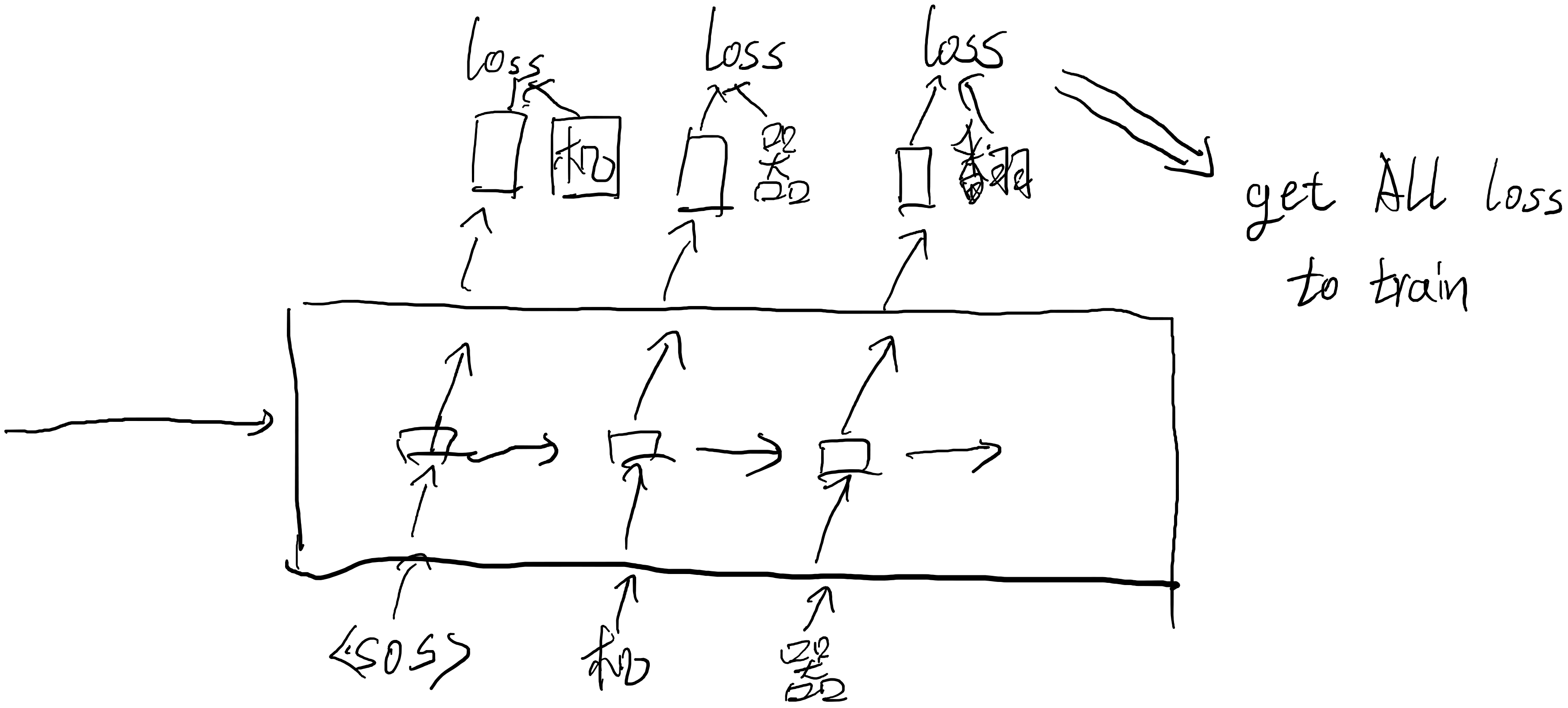
- Solution : Oracle select in decoding
  - World-level oracle selection
  - Sentence-level oracle selection

- Stumbling
  - How to get same sentence length → Force Decoding
  - Real Training → Sampling with Decay

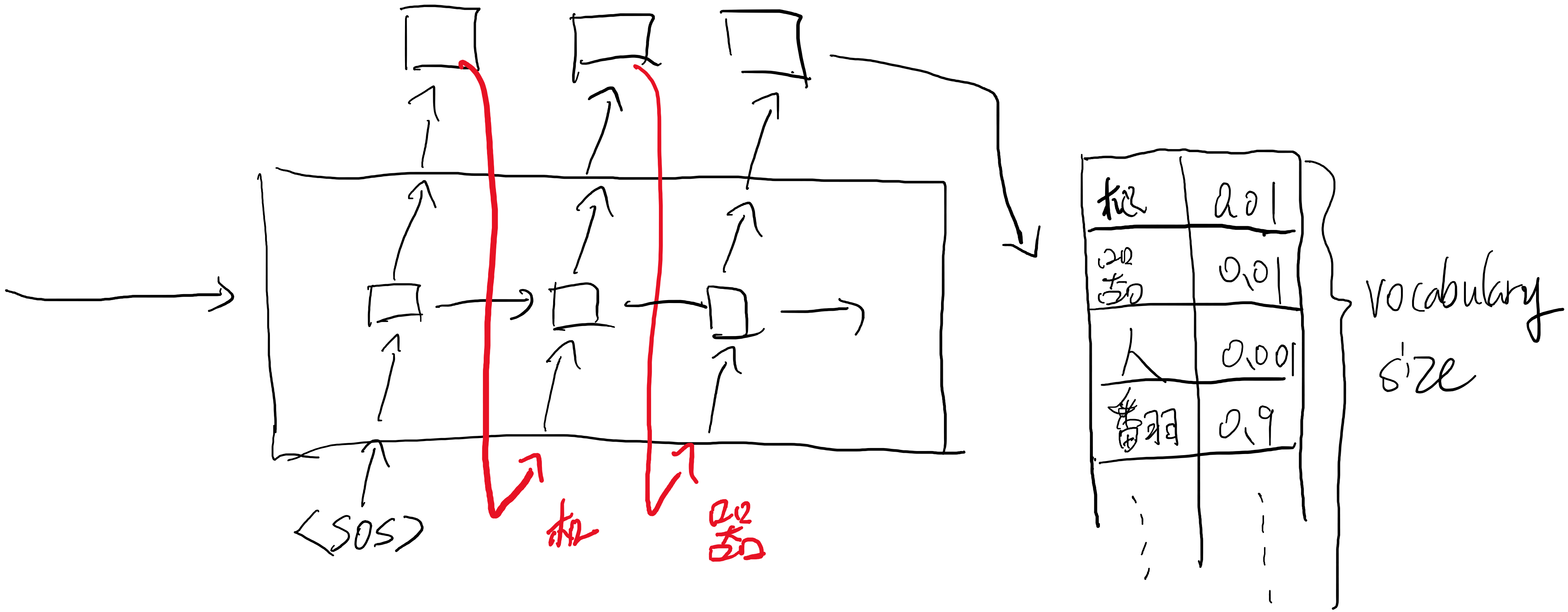
Before the paper



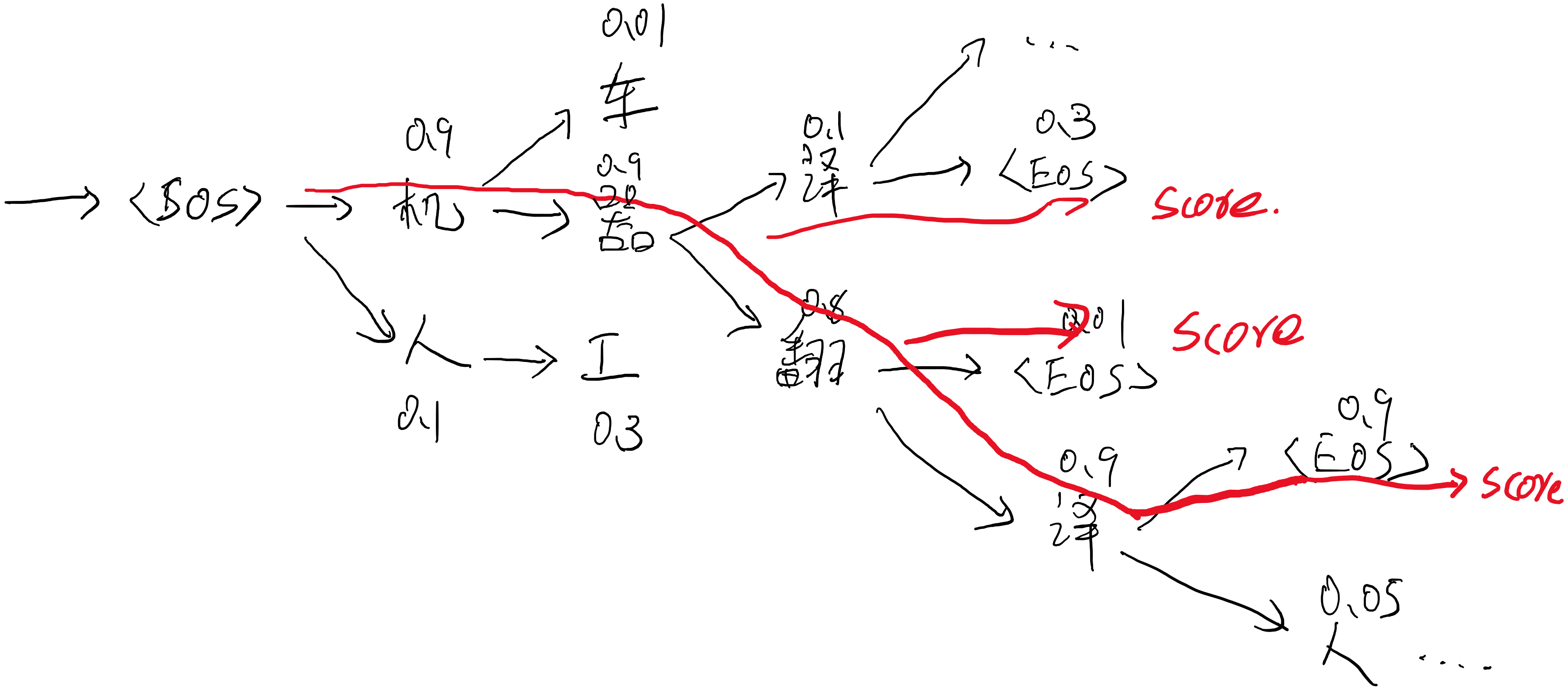
Before the paper



Before the paper



Before the paper



Training and Inference are different

At training time the **ground truth** words are used as context while at **inference the entire sequence is generated by the resulting model on its own and hence the previous words generated by the model are fed as context**.

As a result, the predicted words at training and inference **are drawn from different distributions**, namely, from the data distribution as opposed to the model distribution. This discrepancy, called *exposure bias* (Ranzato et al., 2015), **leads to a gap between training and inference**. As the target sequence grows, the errors accumulate among the sequence and the model has to predict under the condition it has never met at training time.

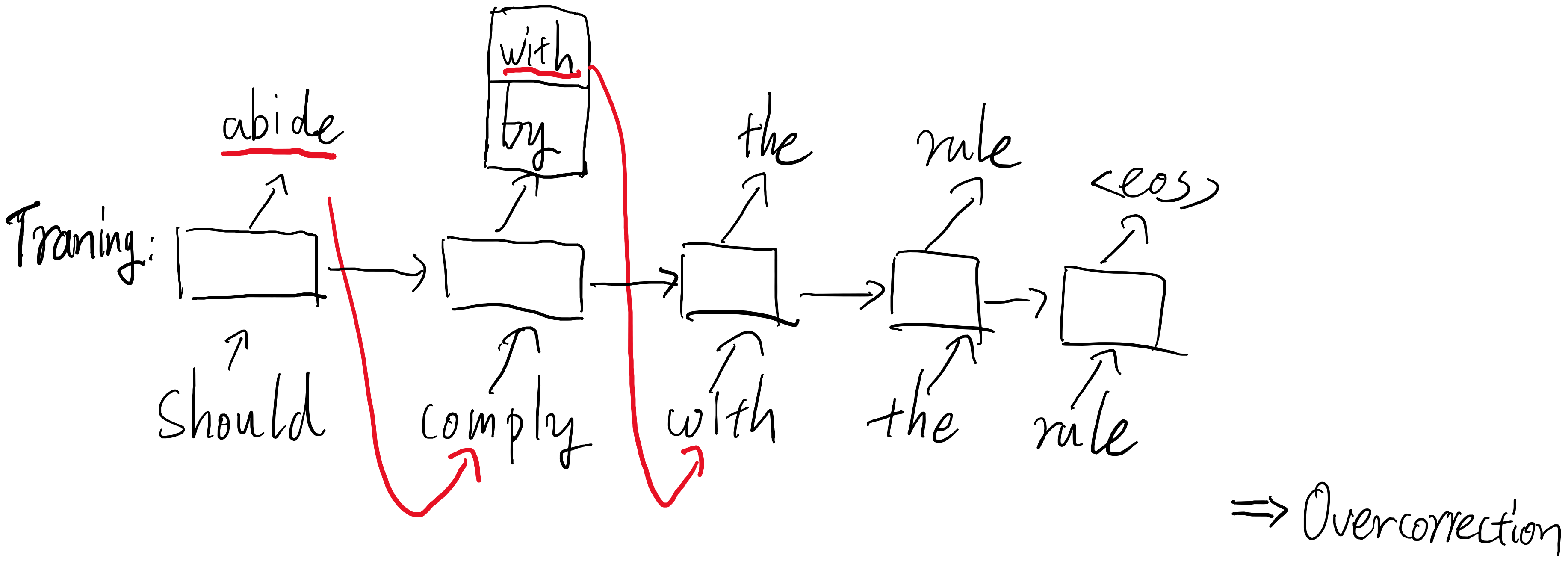
Bad Teacher Forcing

Reference: We should comply with the rule.

Inference: We should abide with the rule.

We should abide by the law.

We should abide by the rule.



Inference: should abide with the rule

Bridging the Gap

- Using non ground true word or sentence in training time : Oracle Word Selection
  - word-level selection
  - Sentence-level selection

- Using Oracle Word Selection in proper way : Sampling with Decay

Bridging the Gap : word-level selection

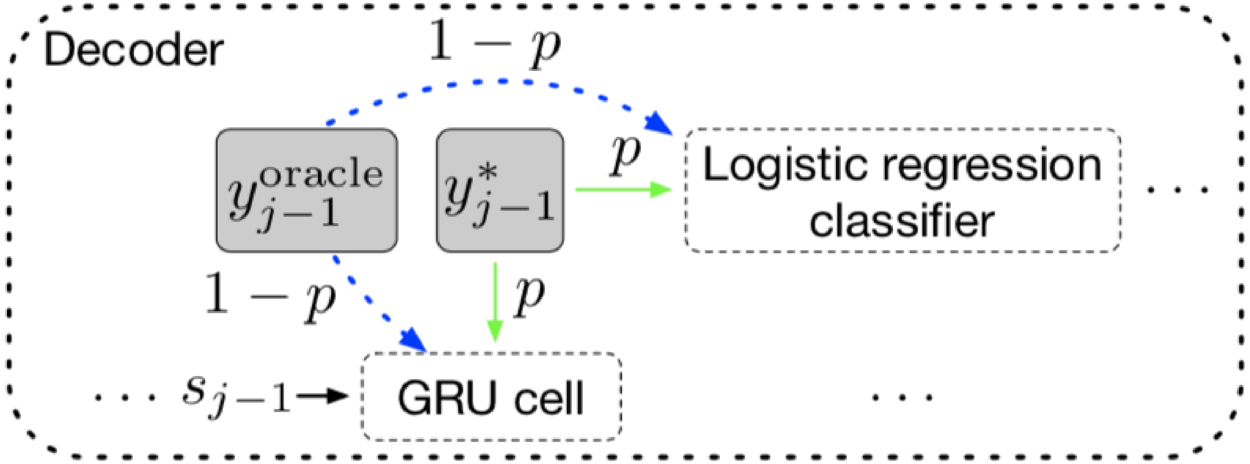
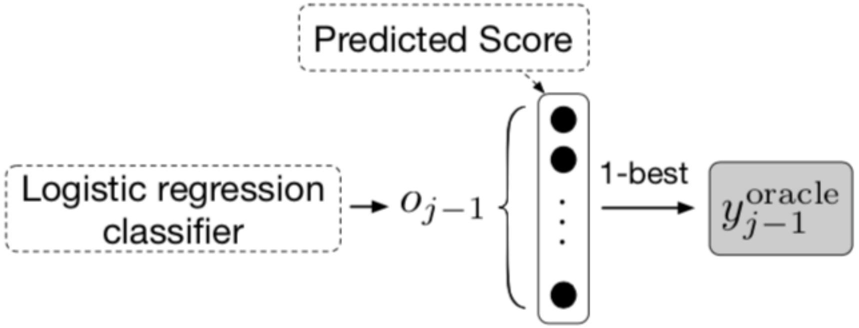
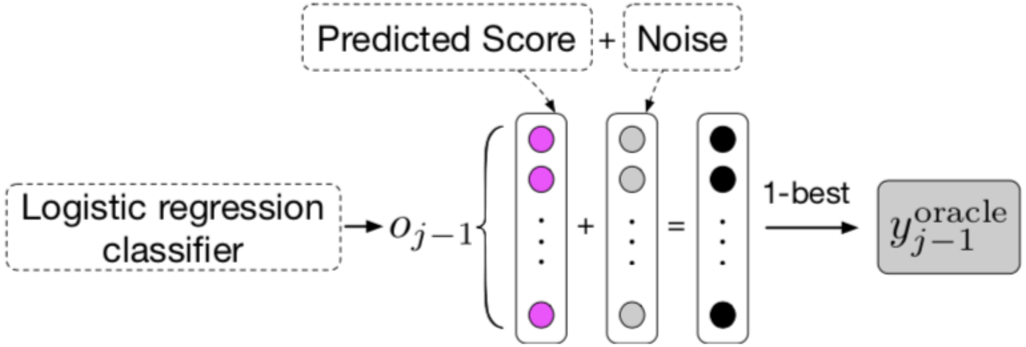


Figure 1: The architecture of our method.





Bridging the Gap : sentence-level selection

Employ **BLEU** as the sentence-level metric. To select the sentence-level oracles, first perform beam search for all sentences in each batch, assuming beam size is k, and **get k-best candidate** translations. In the process of beam search, we also could apply the **Gumbel noise** for each word generation. We then evaluate each translation by calculating its BLEU score with the ground truth sequence, and use the translation with the highest BLEU score as the *oracle sentence*.

Bridging the Gap : sentence-level selection : problem

Problem comes with sentence-level oracle: have different sentence length.

Force Decoding:

As the length of the ground truth sequence is $|y^*|$, the goal of force decoding is to generate a sequence with $|y^*|$ words followed by a EOS symbol. Therefore, in beam search, once a candidate translation tends to end with EOS when it is shorter or longer than $|y^*|$, we will force it to generate $|y^*|$ words:

1. If the candidate translation gets a word distribution P_j at the j-th step where j $|y^*|$ and EOS is the top first word in P_j , then we select the top second word in P_j as the j-th word of this candidate translation.
2. If the candidate translation gets a word distribution $P_{|y^*|+1}$ at the $\{|y^*|+1\}$-th step where EOS is not the top first word in $P_{|y^*|+1}$, then we select EOS as the $\{|y^*|+1\}$-th word of this candidate translation.

Bridging the Gap : Sampling with Decay

- Using the oracle word randomly
- Not use the oracle word at the beginning
- Increasing the oracle word's probability with the training

$$p = \frac{\mu}{\mu + \exp\left(e/\mu\right)}$$

e is epoch

μ is a hyper-parameter

Experience and result ：score

Systems	Architecture	MT03	MT04	MT05	MT06	Average
Existing end-to-end NMT systems						
Tu et al. (2016)	Coverage	33.69	38.05	35.01	34.83	35.40
Shen et al. (2016)	MRT	37.41	39.87	37.45	36.80	37.88
Zhang et al. (2017)	Distortion	37.93	40.40	36.81	35.77	37.73
Our end-to-end NMT systems						
this work	RNNsearch	37.93	40.53	36.65	35.80	37.73
	+ SS-NMT	38.82	41.68	37.28	37.98	38.94
	+ MIXER	38.70	40.81	37.59	38.38	38.87
	+ OR-NMT	**40.40**‡†*	**42.63**‡†*	**38.87**‡†*	**38.44**‡	**40.09**
	Transformer	46.89	47.88	47.40	46.66	47.21
	+ word oracle	47.42	48.34	47.89	47.34	47.75
	+ sentence oracle	**48.31***	**49.40***	**48.72***	**48.45***	**48.72**

Table 1: Case-insensitive BLEU scores (%) on Zh→En translation task. "‡", "†", "⋆" and "∗" indicate statistically significant difference ($p<0.01$) from RNNsearch, SS-NMT, MIXER and Transformer, respectively.

SS-NMT: Our implementation of the scheduled sampling (SS) method (Bengio et al., 2015) on the basis of the RNNsearch. The decay scheme is the same as Equation 15 in our approach.

MIXER: Our implementation of the mixed incremental cross-entropy reinforce (Ranzato et al., 2015), where the sentence-level metric is BLEU and the average reward is acquired according to its offline method with a 1-layer linear regressor.

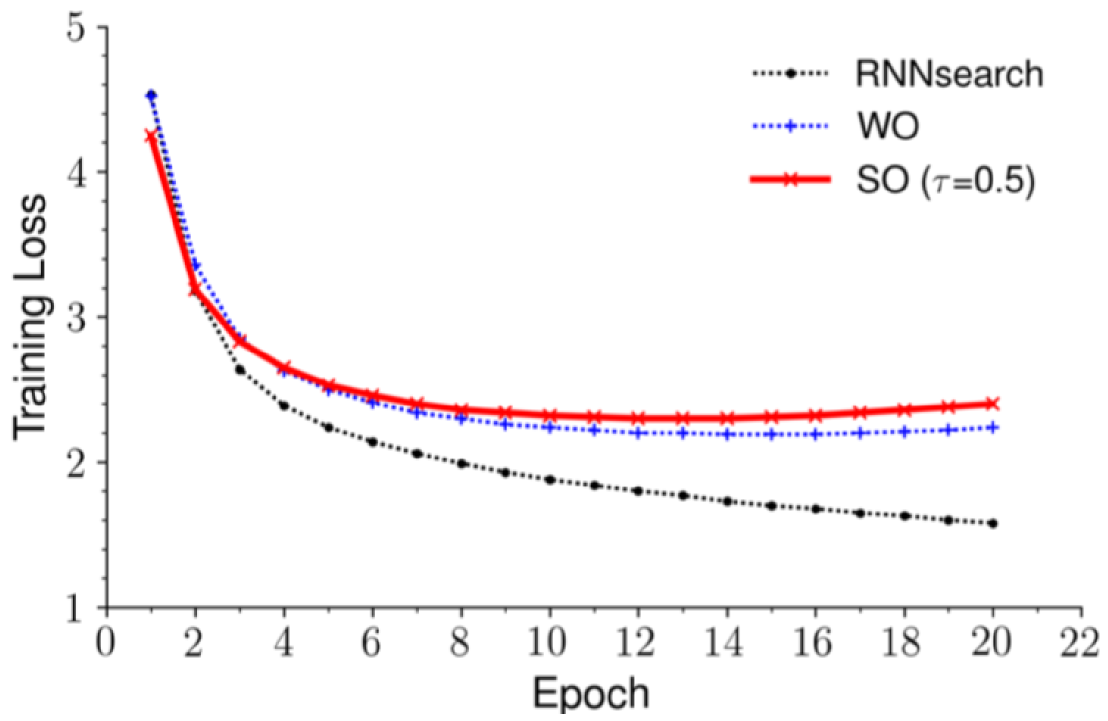
OR-NMT: Based on the RNNsearch, we introduced the word-level oracles, sentence-level oracles and the Gumbel noises to enhance the overcorrection recovery capacity. For the sentence-level oracle selection, we set the beam size to be 3, set τ=0.5 in Equation (11) and μ=12 for the decay function in Equation (15). OR-NMT is the abbreviation of NMT with Overcorrection Recovery.

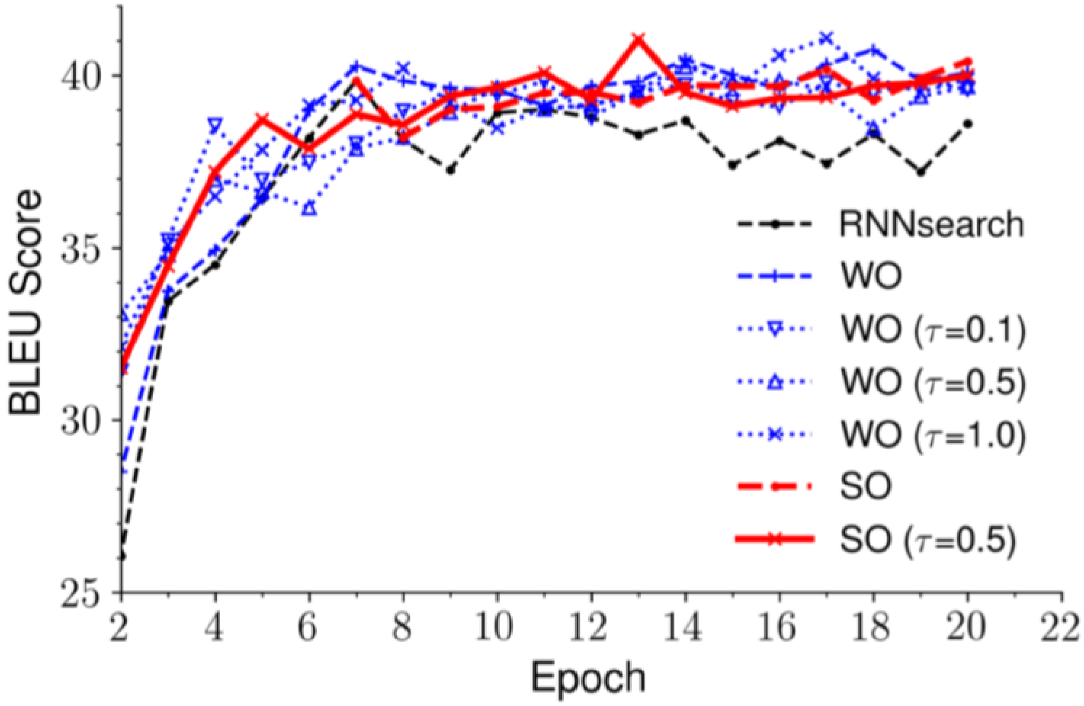
Experience and result ：affect

Systems	**Average**
RNNsearch	37.73
+ word oracle	38.94
+ noise	39.50
+ sentence oracle	39.56
+ noise	**40.09**

Table 2: Factor analysis on Zh→En translation, the results are average BLEU scores on MT03∼06 datasets.

Experience and result ：converge





Comments

- Good:
  - Simple ,sharp and easy to read
  - Great angle to think about the decoding

- Flaw:
  - Using BLEU in sentence oracle is like a trick. Make BELU to be the loss almost
  - Sampling with Decay and word-level oracle more like a normalization as said in this paper
  - BLEU is not a good metric . It’s not a good way to use BLEU to be the oracle word selection.