A Mixed Learning Objective for Neural Machine Translation

Wenjie Lu, Leiying Zhou, Gongshen Liu* and Quanhai Zhang*
School of Electronic Information and Electrical Engineering,
Shanghai Jiao Tong University, Shanghai, China
{jonsey,zhouleiying,lgshen,qhzhang}@sjtu.edu.cn

Abstract

Evaluation discrepancy and overcorrection phenomenon are two common problems in neural machine translation (NMT). NMT models are generally trained with word-level learning objective, but evaluated by sentence-level metrics. Moreover, the cross-entropy loss function discourages model to generate synonymous predictions and overcorrect them to ground truth words. To address these two drawbacks, we adopt multi-task learning and propose a mixed learning objective (MLO) which combines the strength of word-level and sentence-level evaluation without modifying model structure. At word-level, it calculates semantic similarity between predicted and ground truth words. At sentence-level, it computes probabilistic n-gram matching scores of generated translations. We also combine a loss-sensitive scheduled sampling decoding strategy with MLO to explore its extensibility. Experimental results on IWSLT 2016 German-English and WMT 2019 English-Chinese datasets demonstrate that our methodology can significantly promote translation quality. The ablation study shows that both word-level and sentence-level learning objective can improve BLEU scores. Furthermore, MLO is consistent with state-of-the-art scheduled sampling methods and can achieve further promotion.

1 Introduction

In recent years, tremendous progresses have been made in the field of neural machine translation (NMT) (Sutskever et al., 2014; Luong et al., 2015). A typical NMT model can be formulated as an encoder-decoder-attention architecture (Forcada and Ñeco, 1997; Bahdanau et al., 2015) with maximum likelihood estimation (MLE) objective. Given sufficient parallel corpora, NMT models can achieve promising performance.

Despite much success, NMT models suffer from two major drawbacks. First, there exists a discrepancy between training objectives and evaluation metrics. Most NMT models are trained with MLE objective under the teacher forcing algorithm (Williams and Zipser, 1989), i.e., models calculate and accumulate cross-entropy loss between predicted and ground truth sentences word by word. A lower cross-entropy value means the predictions are closer to ground truth at word level. Model parameters are updated through backpropagation to minimize the value of loss function. However, translation quality is measured by sentence-level metrics such as BLEU (Shterionov et al., 2017), ROUGE (Lin, 2004), etc. This way of word-level optimization mismatches sentence-level evaluation metrics, which may mislead the promotion of translation performance. Second, the MLE training objective brings about overcorrection phenomenon (Zhang et al., 2019). To be specific, models are trained to learn absolutely correct translations and overcorrect synonymous words and phrases. Once the model predicts a word different from the ground truth word, the cross-entropy loss will immediately punish it and lead the model to the correct direction. As for synonymous phrases, it may result in translating wrong phrases while reducing the diversity of translation.

In this paper, we present a novel approach to solve the above problems. Instead of training NMT models with word-level cross-entropy loss, we propose to train models with a mixed learning objective

Proceedings of the 19th China National Conference on Computational Linguistics, pages 975-984, Hainan, China, October 31 - November 1, 2020.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

(MLO), which can combine the strength of word-level and sentence-level training. At word level, MLO estimates semantic similarity between the predicted and the ground truth words. Synonymous words will be encouraged rather than overcorrected. At sequence level, MLO calculates probabilistic n-gram matching score between the predicted and the ground truth sentences. The differentiable property of MLO enables NMT models to be trained flexibly without modifying structure. Most important of all, it can relieve the problem of evaluation discrepancy and overcorrection phenomenon.

The major contributions of this paper are summarized as follows:

- We present a novel mixed learning objective for training NMT models, aiming at alleviating evaluation discrepancy and overcorrection phenomenon. The mixed learning objective can encourage word-level semantic similarity and balance sequence-level n-gram precision of the translation.
- We explore the extensibility of mixed learning objective and adopt a novel loss-sensitive scheduled sampling instead of teacher forcing algorithm. The proposed objective is more consistent with state-of-the-art scheduled sampling methods and can achieve better performance.
- We demonstrate the effectiveness of our approach on IWSLT 2016 German-English and WMT 2019 English-Chinese datasets, and achieve significant improvements. Moreover, the mixed learning objective can be flexibly applied by various model structures and algorithms.

2 Related Work

2.1 Evaluation discrepancy

To tackle the problem of discrepancy between word-level MLE objective and sentence-level evaluation metrics, some researches utilize techniques like generative adversarial network (GAN) (Goodfellow et al., 2014) or reinforcement learning (RL) (Sutton et al., 1998). Borrowed idea from DAD (Venkatraman et al., 2015) and beam search (Sutskever et al., 2014; Rush et al., 2015), Ranzato et al. (2015) proposed Mixed Incremental Cross-Entropy Reinforce (MIXER) to directly optimized model parameters with respect to the metric used at inference time. Further, Shen et al. (2016) presented minimum risk training (MRT) to minimize the expected loss (i.e., risk) on the training data. Wieting et al. (2019) proposed to train NMT models with semantic similarity based on MRT. Wiseman and Rush (2016) introduced beam-search optimization schedule for model to learn global sequence scores. Moreover, Lin et al. (2017) proposed RankGAN which can analyze and rank sentences by giving a reference group, and thus achieve high-quality language descriptions.

2.2 Overcorrection Phenomenon

As for overcorrection phenomenon, especially synonymous phrases, one solution is to utilize the model's previous predictions as input in training. The generation inconsistency between training and inference which called exposure bias (Zhang et al., 2019) causes models to overcorrect from synonymous translations and generate wrong phrases. Bengio et al. (2015) firstly proposed a scheduled sampling strategy based on an algorithm called Data As Demonstrator (DAD) (Venkatraman et al., 2015). At every decoding step, a dynamic probability p is used to decide whether to sample from ground truth or the previous word predicted by the model itself. Inspired by their method, Zhang et al. (2019) came up with sampling from ground truth and inferred sentences word by word through force decoding.

3 Methodology

3.1 Model Overview

Without loss of generality, we utilize a common RNN attention model (Bahdanau et al., 2015) as baseline to demonstrate our approach. Suppose that the source sentence $X = (x_1, x_2, ..., x_{T_x})$ and the target sentence $Y = (y_1, y_1, ..., y_{T_y})$. The RNN model encodes the source sentence as follows:

$$h_t = \phi(h_{t-1}, x_t) \tag{1}$$

Proceedings of the 19th China National Conference on Computational Linguistics, pages 975-984, Hainan, China, October 31 - Novermber 1, 2020.

where h_0 is an initial vector and ϕ is a nonlinear function. Then context vector $c_i, i = 1, 2, ..., T_y$ is calculated by:

$$c_i = \sum_{j=1}^{T_x} \alpha_{ij} \cdot h_j \tag{2}$$

where α_{ij} is the attention weight between c_i and h_j.

When the decoder receives the context c_t, it calculates the hidden layer vector s_t by:

$$s_t = f(s_{t-1}, y_{t-1}, c_t) \tag{3}$$

where s_0 is an initial vector, f is a nonlinear function of hidden layers, y_{t-1} is the historical output at time $t-1$ in inference and ground truth word in training, and y_0 is the end flag of source sentence X.

According to the hidden layer state s_t, the probability of inferring the word y_t can be computed by:

$$P(y_t) = softmax(W_o \cdot p(y_t \mid y_1, ..., y_{t-1}, x)) \tag{4}$$

$$p(y_t \mid y_1, ..., y_{t-1}, x) = g(y_{t-1}, s_t, c_t) \tag{5}$$

where g is a nonlinear function and W_o is a mapping matrix.

Finally, given a set of sequence pairs $(X_i, Y_i), i = 1, 2, ..., N$ in the parallel corpora, the training objective is to maximize the likelihood as follows:

$$\hat{\theta}_{MLE} = argmax\{L(\theta)\} \tag{6}$$

where $L(\theta)$ is the loss function computed by:

$$L(\theta) = \sum_{i=1}^{N} logP(Y_i \mid X_i, \theta) = \sum_{i=1}^{N} \sum_{t=1}^{T_y} logP(y_t) \tag{7}$$

3.2 Word-level Semantic Similarity Objective

The original cross-entropy loss measures the probability of predicting right translation for each word, which means it only cares about how to generate ground truth words with maximum likelihood. This may cause two problems. First, generating any other words is discouraged. Although synonymous translations are right in the subjective sense, they will be punished and corrected to ground truth words. Second, suppose that the word with maximum probability is not ground truth word, and the model will choose it as predicted translation. The calculation of cross-entropy loss does not take into consideration of what exactly that word is, which is important for evaluating the model.

Therefore, we design the word-level learning objective in order to measure the semantic similarity between the generated translations and ground truth sentences. There have been lots of complex researches on semantic similarity (Pradhan et al., 2015; Kenter and De Rijke, 2015). In order not to include additional models, we adopt the cosine similarity method for measurement. Mathematically speaking, cosine similarity calculates the semantic similarity between two non-zero vectors, which is suitable for word embeddings.

Given the predicted translation $Y^* = (y_1^*, y_2^*, \ldots, y_{T_{Y^*}}^*)$, the semantic similarity between sentence Y and Y^* can be calculated by:

$$Sim(Y, Y^*) = \sum_{i=1}^{T_y} \frac{emb(y_i) \cdot emb(y_i^*)}{\|emb(y_i)\| \times \|emb(y_i^*)\|} \tag{8}$$

where $emb(\cdot)$ refers to the word embedding of each word.

Therefore, we can calculate semantic similarity between every translation and corresponding ground truth sentence. During training, the word-level training objective is defined as followings:

$$L_{word} = -\sum_{j=1}^{N} Sim(Y_j, Y_j^*) \tag{9}$$

3.3 Sentence-level Probabilistic N-gram Objective

The word-level semantic similarity objective helps to foster translation diversity and relieve the problem of overcorrection, which can improve word-level translation accuracy. As for another important standard fluency in machine translation, we design a sentence-level probabilistic n-gram objective which is consistent with evaluation metrics.

The calculation of n-gram matching is widely used in machine translation evaluation metrics. Take BLEU for example, firstly n-grams in source sentence Y and Y^* are extracted and counted, denoted as $C(n-gram)$. Next, n-gram matches between Y and Y^* are computed and denoted as $C_{clip}(n-gram)$. The precision score can be calculated by their ratio.

However, the non-differentiable property of BLEU makes it unable to be adopted as loss function. Therefore, inspired by Shao et al. (2018), we modified the calculation of n-gram matches as follows. Supposing that $(g_1, g_2, ..., g_n)$ is an n-gram sequence in Y, then its occurrences can be computed by:

$$\widetilde{C}_Y(n-gram) = \sum_{i=0}^{T_Y-n} \prod_{j=1}^{n} 1\{g_j = y_{i+j}\} \cdot P(y_{i+j}) \tag{10}$$

where $1\{\cdot\}$ denotes an indicator function and $P(\cdot)$ is calculated by equation (5). Then, the clip n-gram matches between two sentences and the precision score of translation Y can be computed as follows:

$$C_{clip}(n-gram) = min\{\widetilde{C}_Y(n-gram), C_{Y^*}(n-gram)\} \tag{11}$$

$$\widetilde{p_n} = \frac{\sum\limits_{n-gram \in Y} C_{clip}(n-gram)}{\sum\limits_{n-gram' \in Y} \widetilde{C}_Y(n-gram')} \tag{12}$$

Finally, to punish very long or short translations, BLEU is modified based on $\widetilde{p_n}$ and defined as follows:

$$\widetilde{BLEU}(Y, Y^*) = BP \cdot exp(\sum_{n=1}^{N} w_n log\widetilde{p_n}) \tag{13}$$

where BP is brevity penalty, w_n is positive weights and N is the maximum length of n-gram.

Therefore, we can calculate probabilistic n-gram matching score between every translation and corresponding ground truth sentence. During training, the sentence-level training objective is defined as followings:

$$L_{sent} = -\sum_{j=1}^{N} \widetilde{BLEU}(Y_j, Y_j^*) \tag{14}$$

3.4 Mixed Learning Objective

In order to alleviate the problem of evaluation discrepancy and overcorrection phenomenon, we propose the mixed learning objective. At word level, it can calculate semantic similarity for training evaluation and promote translation diversity. At sentence level, it can compute probabilistic n-gram precision of predicted sentence and promote translation fluency. The mixed learning objective is defined as follows:

$$L_{total} = L_{ce} + \alpha_{word} \cdot L_{word} + \alpha_{sent} \cdot L_{sent} \tag{15}$$

where L_{ce} refers to cross-entropy loss, α_{word} is the weight of word-level loss function and α_{sent} is the weight of sentence-level loss function.

Similar to Zhang et al. (2019), we adopt the Gumbel-Max technique (Gumbel, 1954; Maddison et al., 2014) for generating more robust outputs. To be specific, the Gumbel noise is defined as follows:

$$G = -log(-logU) \tag{16}$$

where $U \sim Unif[0, 1]$.

Then equation (5) is modified to:

$$P(y_t) = softmax(\frac{W_o \cdot p(y_t \mid y_1, ..., y_{t-1}, x) + G}{\tau}) \tag{17}$$

where τ is a temperature parameter controlling the generated distribution.

During training, we adopt a scheduled sampling strategy instead of teacher forcing algorithms. At every decoding step, a probability p is used to decide whether to sample from ground truth or inferred words. Specifically, assuming w_t is the input at each decoding step t and y'_{t-1} is the word obtained from inferred words, then $Pr(w_t = y_{t-1}) = p$ and $Pr(w_t = y'_{t-1}) = 1 - p$. We hope the probability p to decay from 1 to 0, so that the training process can gradually learned to deal with simulated inference situation.

Borrowing idea from the decay schedule in learning rate, sample probability can be defined as an inverse sigmoid curve with variable training epochs. Considering that a loss function intuitively reflects how well the model is trained, we define loss-sensitive sample probability as follows:

$$p = \frac{k}{k + exp(\frac{e}{k})} \cdot \sigma(L) \tag{18}$$

where k is a hyper-parameter, e is the current index of epoch, L is the average loss function value of epoch e, and σ is a non-linear function. For practice, we choose tanh function.

4 Experiments

4.1 Experimental Setup

We conduct our experiments comparable with previous work by using the following two datasets:

German-English. The German-English dataset is chosen from IWSLT 2016 (Cettolo et al., 2012). We use official testset2013 as validation set. The training and validation data consists of 196,884 and 992 sentences respectively. As for evaluation, we use the testset dataset from 2010 to 2014 and tokenized BLEU scores as computed by the multi-bleu.perl script[0].

English-Chinese. The English-Chinese dataset is chosen from the casia2015 parallel corpus in WMT 2019 shared task. It consists of approximately 1.05M sentences. We use official newsdev2017 as validation set and evaluate on the newstest dataset from 2017 to 2019.

For all training data, we perform tokenization and truecasing using standard Moses tools. For Chinese corpora, we use jieba[1] for segmentation. Then, we employ byte pair encoding (BPE) (Sennrich et al., 2016) with 50,000 operations to alleviate Out-of-Vocabulary problem. To accelerate training and save cost, we discard sentences with more than 50 tokens. The dimension of word embeddings is set to 512.

We first pretrain the baseline model by MLE. Then, we replace the cross-entropy loss function with MLO. The model is trained with a batch size of 60. We use Adam (Kingma and Ba, 2014) optimizer to tune the parameters. Besides, we use dropout regularization with a drop probability 0.5. During decoding, the beam size is set to 3. The hyper-parameter of sample probability k and temperature τ are set to 12 and 0.5 respectively. The weights of word-level and sentence-level loss function are set to 0.8 and 150 respectively.

4.2 Baseline Systems

We compare our method with existing common NMT systems including Transformer(Vaswani et al., 2017), Evolved Transformer (So et al., 2019) and DTMT (Meng and Zhang, 2019). Moreover, to explore the extensibility of MLO, we experiment on several state-of-the-art scheduled sampling works. These baseline systems are included as follows:

RNNsearch. A vanilla attention-based recurrent neural network which consists of 2-layer bidirectional GRU units (Cho et al., 2014). The dimension of hidden layer is 512.

[0] https://github.com/moses-smt/mosesdecoder/blob/master/scripts/generic/multi-bleu.perl

[1] https://github.com/fxsjy/jieba

Proceedings of the 19th China National Conference on Computational Linguistics, pages 975-984, Hainan, China, October 31 - November 1, 2020.

SS-NMT. A word-level scheduled sampling method (Bengio et al., 2015) which utilizes an inverse sigmoid decay schedule to sample from previous predicted word and ground truth word.

OR-NMT. A sentence-level scheduled sampling method (Zhang et al., 2019) which utilizes inverse sigmoid decay schedule to sample from predicted sentence and ground truth sentence. Predicted sentence is generated by beam search and force decoding.

4.3 Main Results

Table 1: Results of the proposed method on German-English dataset (BLEU).

Systems	testset2010	testset2011	testset2012	testset2014	average
Transformer	25.17	30.03	26.20	24.24	26.41
Evolved Transformer	26.33	31.45	27.28	25.36	27.61
DTMT	26.51	31.66	27.64	26.02	27.96
RNNsearch	24.46	28.06	24.92	22.94	25.10
+ SS-NMT	26.46	30.14	26.60	24.31	26.88
+ OR-NMT	27.37	30.72	27.54	25.20	27.71
+ MLO	25.84	29.85	26.32	23.73	26.44
+ SS-NMT + MLO	26.78	30.34	26.99	24.81	27.23
+ OR-NMT + MLO	**27.44**	**31.89**	**27.65**	**25.92**	**28.22**

Table 2: Results of the proposed method on English-Chinese dataset (BLEU).

Systems	newstest2017	newstest2018	newstest2019	average
Transformer	26.37	25.09	25.76	25.74
Evolved Transformer	27.84	25.98	27.25	27.02
DTMT	28.07	26.10	27.34	27.17
RNNsearch	24.92	24.17	24.20	24.63
+ SS-NMT	25.89	25.12	25.43	25.48
+ OR-NMT	28.03	26.10	26.66	26.93
+ MLO	25.83	24.74	25.32	25.29
+ SS-NMT + MLO	26.60	25.42	25.63	25.88
+ OR-NMT + MLO	**28.18**	**26.63**	**27.13**	**27.31**

TABLE 1 and TABLE 2 reports the results of the proposed method in comparison to other NMT systems on German-English and English-Chinese datasets respectively. As it can be seen, training with sentence-level scheduled sampling and mixed learning objective (OR-NMT + MLO) obtains the best published results on all testsets.

On German-English dataset, our full system can outperform RNNsearch by +3.11 BLEU averagely. On English-Chinese dataset, our full system can have an improvement of +2.68 BLEU on three testsets.

To validate the effectiveness of mixed learning objective, we carry out ablation study to evaluate the performance of word-level and sentence-level learning objective respectively. The mixed learning objective is proposed to encourage word-level semantic similarity and balance sequence-level n-gram precision of the translation. Meanwhile, it aims at relieving the problem of evaluation discrepancy and overcorrection. We will display and analyze the effect of mixed learning objective in detail in Section 4.4.

Another point of focus lies in the extensibility of mixed learning objective. As shown in TABLE 1 and TABLE 2, combining scheduled sampling strategy with mixed learning objective can achieve better translation performance. We will discuss the effect of scheduled sampling from two aspects in Section

Table 3: BLEU scores on German-English dataset.

Systems	testset2010	testset2011	testset2012	testset2014	average
RNNsearch	24.46	28.06	24.92	22.94	25.10
+ L_{word}	25.18	28.57	25.74	23.83	25.83
+ L_{sent}	25.40	28.72	26.01	24.07	26.05
+ MLO	25.84	29.85	26.32	23.73	26.44

Table 4: BLEU scores on English-Chinese dataset.

Systems	newstest2017	newstest2018	newstest2019	average
RNNsearch	24.92	24.17	24.20	24.63
+ L_{word}	25.26	24.45	24.67	24.79
+ L_{sent}	25.59	24.51	25.10	25.06
+ MLO	25.83	24.74	25.32	25.29

4.5. Besides, the loss-sensitive sample probability is defined to sense the speed of converge and make adjustment on sample probability. We will analyse its effect on scheduled sampling methods to explore how to achieve better performance.

4.4 Effect of Mixed Learning Objective

Aiming to alleviate evaluation discrepancy and overcorrection phenomenon, we propose the mixed learning objective which can promote word-level semantic similarity and sequence-level n-gram precision. To explore the effect of mixed learning objective, we conduct experiments on word-level and sentence-level learning objective respectively without scheduled sampling strategy on RNNsearch under the same conditions.

The experimental results are listed in TABLE 3 and TABLE 4. As it can be seen, only using word-level or sentence-level learning objective rather than cross-entropy loss can help achieve higher BLEU scores on two datasets. To be specific, word-level learning objective can get a promotion of +0.16 $\sim$ +0.73 BLEU averagely over RNNsearch on German-English and English-Chinese datasets. Sentence-level learning objective can outperform RNNsearch by +0.43 $\sim$ +0.95 BLEU score on two datasets averagely.

For the experimental results, we make some simple analysis. The word-level learning objective takes into account semantic similarity between predicted and ground truth words, so that it can avoid forcing model to generate the only one correct translation. The promotion in BLEU scores verifies that discouraging and punishing other synonymous words is disadvantageous for NMT models. Therefore, the word-level learning objective can to some extent solve this problem and encourage translation diversity.

The sentence-level learning objective calculates probabilistic n-gram matching scores between predicted and ground truth sentence, which is coordinate with general evaluation metrics. On the one hand, it contributes to alleviate the problem of evaluation discrepancy without importing additional complex model. On the other hand, the objective can naturally promote translation performance on BLEU scores.

Furthermore, MLO which combines word-level and sentence-level learning objective can obtain best translation performance in BLEU scores. Specifically, MLO can outperform RNNsearch by +0.66 $\sim$ +1.34 BLEU score averagely on German-English and English-Chinese datasets.

4.5 Effect of Loss-sensitive Scheduled Sampling

To validate the extensibility of MLO, we conduct various experiments which combine MLO with state-of-the-art scheduled sampling methods. Moreover, we defined a novel loss-sensitive sample probability for model to be flexibly adapted to scheduled sampling strategy. Under the same experimental settings, we conduct experiments on German-English and English-Chinese datasets to validate the effectiveness of loss-sensitive scheduled sampling and analyze in two aspects.

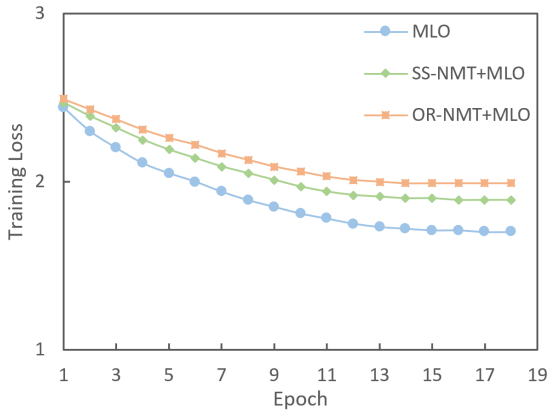

Figure 1: The training loss curves of three baseline systems on the IWSLT 2016 German-English translation task.

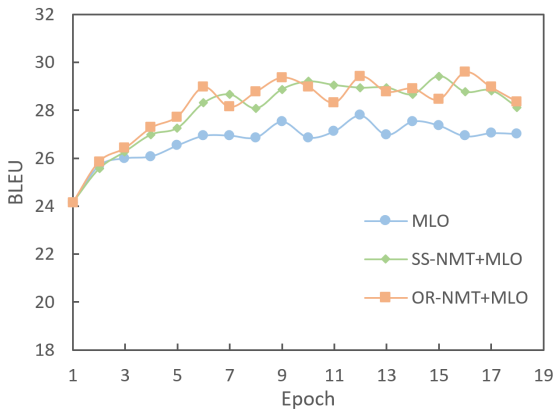

Figure 2: Trends of BLEU scores of three baseline systems on the validation set on the IWSLT 2016 German-English translation task.

Fig. 1 gives the training loss curves of MLO, SS-NMT+MLO and OR-NMT+MLO during training. As the training epoch increases, MLO continues to decrease at the lowest value and gradually tends to be flat. Due to different sampling strategies, SS-NMT and OR-NMT gradually converge to a certain training loss value. Moreover, Fig. 2 gives the BLEU score curves of three methods. It can be seen that SS-NMT+MLO and OR-NMT+MLO can achieve better BLEU scores compared to RNNsearch on validation set. We can also conclude from TABLE 1 and TABLE 2 that SS-NMT+MLO can achieve a promotion of +0.35 $\sim$ +0.4 BLEU scores over SS-NMT and OR-NMT+MLO can outperform OR-NMT by +0.38 $\sim$ +0.51 BLEU scores.

Since the starting point of scheduled sampling is to solve the problem of exposure bias and overcorrection phenomenon, the original cross-entropy loss function may be hard to score the inference results and guide the training process. However, the MLO is proposed for alleviating these problems as well. Therefore, the idea of combining MLO with scheduled sampling is natural and proved to be effective.

Besides the mutual promotion of MLO and scheduled sampling, the last thing we want to point out is the necessity and effectiveness of loss-sensitive sample probability. We define $\sigma(L) = 1$ as non-sensitive sample probability and perform parallel tests. By observing their decay curves during training, we find that loss-sensitive sample probability is more flexible and helpful in adjusting a proper probability for different training scenes. Since $tanh(L) < 1$, the loss-sensitive probability is calculated to be lower than non-sensitive probability. From the perspective of feeding as input inferred rather than ground truth words, we make it harder for model to learn and correct mistakes. Meanwhile, the experimental results show promotion on translation quality.

5 Conclusion

In this paper, we propose a mixed learning objective for NMT so as to alleviate the problem of evaluation discrepancy and overcorrection phenomenon. At word-level, the objective measures semantic similarity between the generated and ground truth words. At sentence-level, the objective calculates probabilistic n-gram matching scores of the translations. Moreover, we combine loss-sensitive scheduled sampling methods with mixed learning objective for mutual promotion. Experimental results show that our proposed method can achieve significant improvement on BLEU scores compared to previous works.

Acknowledgment

This research work has been funded by the National Natural Science Foundation of China (Grant No.61772337), the National Key Research and Development Program of China NO. 2018YFC0830803.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In Yoshua Bengio and Yann LeCun, editors, *ICLR*.

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. 2015. Scheduled sampling for sequence prediction with recurrent neural networks. In *NIPS*, pages 1171–1179.

Mauro Cettolo, Christian Girardi, and Marcello Federico. 2012. Wit^3: Web inventory of transcribed and translated talks. In *EAMT*, pages 261–268, Trento, Italy.

Kyunghyun Cho, Bart van Merriënboer Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *EMNLP*, pages 1724–1734.

Mikel L Forcada and Ramón P Ñeco. 1997. Recursive hetero-associative memories for translation. In *IWANN*, pages 453–462. Springer.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *NIPS*, pages 2672–2680.

Emil Julius Gumbel. 1954. Statistical theory of extreme values and some practical applications. *NBS Applied Mathematics Series*, 33.

Tom Kenter and Maarten De Rijke. 2015. Short text similarity with word embeddings. In *CIKM*, pages 1411–1420.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kevin Lin, Dianqi Li, Xiaodong He, Zhengyou Zhang, and Ming-Ting Sun. 2017. Adversarial ranking for language generation. In *NIPS*, pages 3155–3165.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. In *EMNLP*, pages 1412–1421.

Christopher Maddison, Daniel Tarlow, and Tom Minka. 2014. A* sampling. *NIPS*, 10.

Fandong Meng and Jinchao Zhang. 2019. Dtmt: A novel deep transition architecture for neural machine translation. In *AAAI*, volume 33, pages 224–231.

Nitesh Pradhan, Manasi Gyanchandani, and Rajesh Wadhvani. 2015. A review on text similarity technique used in ir and its application. *International Journal of Computer Applications*, 120(9).

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2015. Sequence level training with recurrent neural networks. *arXiv preprint arXiv:1511.06732*.

Alexander M Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. In *EMNLP*, pages 379–389.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *ACL*, pages 1715–1725.

Chenze Shao, Yang Feng, and Xilin Chen. 2018. Greedy search with probabilistic n-gram matching for neural machine translation. *arXiv preprint arXiv:1809.03132*.

Shiqi Shen, Yong Cheng, Zhongjun He, Wei He, Hua Wu, Maosong Sun, and Yang Liu. 2016. Minimum risk training for neural machine translation. In *ACL*, pages 1683–1692.

Dimitar Shterionov, Pat Nagle, Laura Casanellas, Riccardo Superbo, and Tony O'Dowd. 2017. Empirical evaluation of nmt and pbsmt quality for large-scale translation production. In *EAMT: User Track*.

David So, Quoc Le, and Chen Liang. 2019. The evolved transformer. In *ICML*, pages 5877–5886.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *NIPS*, pages 3104–3112.

Proceedings of the 19th China National Conference on Computational Linguistics, pages 975-984, Hainan, China, October 31 - Novermber 1, 2020.

Richard S Sutton, Andrew G Barto, et al. 1998. *Introduction to reinforcement learning*, volume 2. MIT press Cambridge.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NIPS*, pages 5998–6008.

Arun Venkatraman, Martial Hebert, and J Andrew Bagnell. 2015. Improving multi-step prediction of learned time series models. In *AAAI*, pages 3024–3030.

John Wieting, Taylor Berg-Kirkpatrick, Kevin Gimpel, and Graham Neubig. 2019. Beyond bleu: Training neural machine translation with semantic similarity. In *ACL*, pages 4344–4355.

Ronald J Williams and David Zipser. 1989. A learning algorithm for continually running fully recurrent neural networks. *Neural computation*, 1(2):270–280.

Sam Wiseman and Alexander M Rush. 2016. Sequence-to-sequence learning as beam-search optimization. In *EMNLP 2016*, pages 1296–1306.

Wen Zhang, Yang Feng, Fandong Meng, Di You, and Qun Liu. 2019. Bridging the gap between training and inference for neural machine translation. In *ACL*, pages 4334–4343.

Proceedings of the 19th China National Conference on Computational Linguistics, pages 975-984, Hainan, China, October 31 - Novermber 1, 2020.