Sentence-level Objectives for Non-Autoregressive Neural Machine Translation

Manolis Rerres
m.rerres@gmail.com

Isa-Ali Kirca
isaali.kirca@hotmail.com

Abstract

Non -Autoregressive models dominate the field of Neural Machine Translation. We employ three different NA transformer CMLM models with the use of two variations of Reward Based loss functions and we compare our results with an implemented NA baseline model with Cross Entropy loss. In addition we set side by side our BLEU and ChrF scores with an AR baseline model and other similar NA MRT models from the literature.

1 Introduction

One of the most common tasks in the field of NLP, with numerous real life applications, is Machine Translation. Multiple models such as Recursive and Convolutional Neural Networks are able to perform this task, though Transformer based models seem to highly outperform all others. One the first such proposed state of art model architectures consists of the encoder and the decoder network part (Vaswani et al., 2017). The encoder encodes a source sentence X, creates its embeddings on a token level and the decoder decodes the representations and regenerates the translated target sentence in an autoregressive way (Sutskever et al., 2014). The above procedure requires multiple passes of the decoder layer as target tokens that form a target sentence are also used as input to predict the new target tokens in the next iteration. Autoregressive Machine Translation models achieve great performance wise results, though their high running time does not allow them to be successful under settings that require low latency. Non-Auto regressive Machine Translation models assume that target tokens are predicted independently from the previous generated words, by only taking into account the source sentence. Thus, a single decoder pass is enough to generate all target tokens in parallel, vastly reducing required decoding time (Guo et al., 2019). A severe drawback of such models is that as a training objective they compute cross entropy loss function on a token level. In essence, the target output of the model sentence is compared word by word with the source translated sentence. Every mismatch between those two, is penalized severely causing sometimes the model to steer its output towards the wrong direction. Due to language's nature, a common phenomenon is that two sentences might express the same semantic meaning even if they are not verbally identical. Computing cross entropy as described in the aforementioned cases will cause the NAT MT model to output a target sentence with worse translational quality than before. Furthermore, as the translation is corrected in later model iterations still under the above scheme, we often observe the generation of duplicate target tokens. This observation was characterized as an overcorrection error by (Tan et al., 2020). Mainly it is attributed to the fact that model assesses the quality of the generated translation, assuming independence between target translate tokens and not performing a semantic comparison on a sentence level. Although Non - Autoregressive models severely speed up the procedure, they seem to produce worse qualitative results in comparison to Auto Regressive ones.Recent scientific works focus on combining the benefits that each of the aforementioned methods come with. An interesting approach to this matter by (Ghazvininejad et al., 2019), suggests a model that uses a non-autoregressive decoder alongside an autoregressive part. Specifically, they introduce a Mask-predict decoding algorithm (CMLM) that increases the translation quality by iteratively refining target tokens in an autoregressive way. To alleviate issues related to the use of Cross Entropy as loss function, (Shen et al., 2015) propose to optimize directly a sentence level evaluation metric, known as Minimum Risk Taking (MRT). This metric may be computed on a corpus, sentence or even a token level. To tackle problems related to target token independent assumption and increase models performance it is wiser to make use of metrics that produce and compare a score for a single or more sentences. Though using MRT isn't a panacea. Most of the times these metrics are not differentiable therefore, reward based reinforcement learning techniques are adopted to optimize the discrete objective during the training phase (Bahdanau et al., 2014).

We train a Non-Autoregressive translation model with CMLM and the use of MRT as a loss function. We pick BLUE and ChrF as sentence level evaluation metrics that will provide a score based on how good is the translation quality of our predicted sentence. We compare our results with other NA model approaches and settings described more extensively in the Related Work section. As the main baseline we use an NA model trained with cross entropy as a cost function. Our fine tuned CMLM model performs worse than the baseline in terms of both aforementioned metrics.

Our work focuses on whether MRT provides a sense of an autoregressive approach to our NA model and boosts its performance. Which are the most appropriate metrics to be used and under what scenarios. Lastly, an interesting research question would be whether a simultaneous use of various rewards in the loss function of a CMLM MRT model may boost its performance.

2 Related Work

An alternative training objective for NAT models, minimizes the Bag of Ngrams difference between the reference and the target translated sentence (Shao et al., 2021). That way they overcome the issues of exact target token matching and the generation of duplicate predicted words that a typical NAT loss comes with. In addition, such an objective is also easily and cost efficient differentiable Though, in most of the cases it causes the search space to expand exponentially, especially for a large number of N.

(Ghazvininejad et al., 2019), is the first work that proposes the CMLM model in the form that we also employ it for our experiments. They provide an analytical configuration for the transformers hyperparameter fine tuning. Their model achieves about 4 points higher BLEU score on specific dataset than any previous state of art NAT models. Though, they only use BLEU score as a metric and Cross Entropy as a loss function. (Wieting et al., 2019), introduce SIMILE, a continuous sentence level reward function based on semantic similarity. They compare their metric with BLEU and METEOR on two NMT datasets and present their correlation with human annotators.

3 Methodology

3.1 CMLM Model

We employ a conditional masked language model (CMLM) that predicts target tokens (Y_{mask}) given the source sentence X and a subset of Y target tokens that have been predicted in a previous iteration. Y_{mask} tokens are independent of each other. Model outputs the conditional probability for each $y \in Y_{mask}$, $P(y|X, Y_{obs})$. Our model's architecture is a typical encoder-decoder transformer scheme. Its decoder part is bi-directional, that means it exploits both left and right context to make the target prediction. It uses a pre-trained transformer in the standard WMT 2014 English-German dataset consisting of about 4.5 million sentence pairs (Vaswani et al., 2017). Therefore its parameter values are also fine tuned for this specific task, sentence translation from English to German language. In the training procedure, during each iteration Y_{mask} tokens are randomly selected from the Y target tokens of the predicted sentence. Those words are masked and the model predicts them in a parallel way. Masking predictions that we are not confident of their quality, allows us to take into account previous robust token generations and recompute their conditional probabilities with a better insight.

3.2 Reinforcement loss

In general the cross-entropy objective is trained with maximum likelihood estimation (MLE). Only the perfect reference translation gets rewarded in this case. However, in the case of a Reinforcement Loss (RL), the model outputs which would obtain a high (not perfect) reward, get encouraged by rewards as well. These RL objectives are usually not differentiable due to their discrete nature. To find an unbiased estimation of the gradient, the objective is reformulated using the log derivative trick (also referred to as the policy gradient):

$$\nabla_\theta \mathrm{L}_\theta = \sum_{t=1}^{T} \nabla_\theta - \log\left(p_t\left(y_t \mid \mathbf{X}, \theta\right)\right) \cdot r(\mathbf{Y}) \quad (1)$$

First the softmax has been calculated from the output of the model, whereafter multinomial sampling from the softmax has been adopted to calculate the reward ($r(\mathbf{Y})$). Multinomial sampling is better than beam search in reward computation and significantly enhances the NMT model performance (Wu et al., 2018). Hence for this reason, multinomial sampling has been used. This reward is calculated to evaluate the translation quality and weights the log-probability of the sampled sentence, where it acts as the learning rate to learn high-quality samples.

3.3 Sentence-level metrics

Two sentence-level (string-based) metrics are used in this experiments, i.e. BLEU and ChrF. With string based metrics, ChrF is capable of achieving a higher accuracy than the widely used BLEU (Kocmi et al., 2021). To test whether the same is applicable to our experiments, we implement both and compare them with each other.

Bilingual Evaluation Understudy Score (BLEU), firstly introduced by (Papineni et al., 2002) is a metric for evaluating our generated Y target sentence to a reference sentence X. A perfect match between sentences would result in a score of 100. According to (Wieting et al., 2019), it has been in the vast majority of the papers in the field of NMT, as it is quick and inexpensive to compute, intuitive, language independent and correlated highly with human translation evaluation. In its gist, it counts matching ngrams between the translated generated sentence and its reference. Comparison is done without taking token order into account. There isn't enough evidence to characterize $BLEU$ as a biased metric that is not suitable for evaluating NMT (Kocmi et al., 2021). Although pretrained metrics like $COMET$ perform better, they support only a small set of languages and might prove to be biased towards the dataset they are trained with.

ChrF is an $F - score$ metric based on the character n-grams, firstly proposed by (Popović, 2015). As the comparison is performed on a character level, it captures various morpho-syntactic phenomena. Both parameter values that need to be specified for its use, number of ngrams n and $beta$ the weight between Recall and Precision , do play a significant role to its performance.

4 Experiments

In this paper three experiments have been conducted. In all three the experiments a non-autoregressive model (CMLM) has been trained using the framework of fairseq[1]. Firstly, the CMLM has been trained with the provided nat_loss criterion to serve as a baseline (NAR Base). For the second and third experiments the CMLM has been trained with the reinforcement loss explained earlier, using BLEU and ChrF as sentence-level metrics (rewards) respectively. All experiments use the pre-trained baseline model of (Ghazvininejad et al., 2019) and have been trained for two epochs.

4.1 CMLM with reinforcement loss

The second experiment used BLEU as the sentence-level metric to evaluate the quality of the translation, while the third experiment used ChrF as the sentence-level metric. In both experiments we first sampled multinomially from the softmax to get the index of a token. To be able to calculate the sentence-level metrics, we divided both the targets as well as the whole sample with indices into a sample with sentences (a list of lists which consist of the indices belonging to that sentence).

The second experiment (BLEU) comprised of two sub experiments. First, the sentence-level metric was calculated using the Natural Language Toolkit (NLTK) library[2], while the second was calculated using the metric from the torchmetrics[3]. Both metrics expect different forms of input.

In the third experiment (ChrF) the sentence-level metric was calculated using torchmetric[4].

To calculate the final loss these rewards were multiplied with the minus log-probability ($-\log\left(p_t\left(y_t \mid \mathbf{X}, \theta\right)\right)$) belonging to the token in the sentence (same reward for all tokens in the same sentence). In the end we summed all these calculations and divided by the length of the targets to get the final loss.

Table 1, sums up the experimental results of our work. Training of all models was performed for two epochs. We do not perform the training of the AR baseline, but we use it for comparison reasons (Shao et al., 2021). By far the most time consuming training procedure was computing the Torch ChrF2 reward, observation which aligns with the literature.

[1] https://github.com/facebookresearch/fairseq/blob/main/examples/nonautoregressive_translation/README.md#translate

[2] https://www.nltk.org/_modules/nltk/translate/bleu_score.html

[3] https://torchmetrics.readthedocs.io/en/v0.8.0/text/bleu_score.html

[4] https://torchmetrics.readthedocs.io/en/stable/text/chrf_score.html

	Time (min)	BLEU	ChrF
AR Base	-	33.72	-
NAT Base	-	29.35	-
R-Base	-	29.85	-
NAR Base	46	24.42	56.24
NLTK BLEU4	60	0.19	-
Torch BLEU4	60	23.16	55.45
Torch ChrF2	360	20.22	53.41

Table 1: BLEU and ChrF scores on the test data - AR Base, NAT Base and R-Base (reinforcement base) scores taken from (Shao et al., 2021)

It is intuitive as well, as score comparison is made on a character level.

None of our employed NA with an MRT loss function translation models, outperforms AR or even NAR baseline models in terms of BLEU and ChRF scores. The quality of our generated translation sentences may be characterized as 'the gist is clear, but has significant grammatical errors' as far as NAR Base, Torch BLEU4 and Torch ChrF2, BLEU scores are concerned. NLTK BLEU4 completely fails the task, but this may be due to an implementation error. Lastly, ChrF scores for our employed CMLM models with Torch BLEU4 and Torch ChrF2 reward function appear to be really close with NAR base score, but still come short.

	Exp.1	Exp.2	Exp.3	Exp.4
Val. loss 1	4.316	2.534	0.199	0.587
Val. loss 2	4.665	1.553	0.070	0.979

Table 2: Validation losses for the two epochs - Experiment 1: NAR Base, Experiment 2: NLTK BLEU, Experiment 3: Torch BLEU, Experiment 4: Torch ChrF

Table 2, depicts loss values computed on the validation set for both epochs across our experiments. We observe that either its value increases after the first iteration, or it sharply falls and almost reaches zero levels.

5 Conclusion

NAT Base and R-Base, represent CMLM models with a cross entropy and a reinforcement learning reward as their loss function, respectively. Their BLEU scores correspond to the ones reported in (Shao et al., 2021) article. Clearly our Non - Autoregressive Models with the use of Minimum Risk Taking do not reach the levels of translation quality that may be found in the literature. Table 2 shows numerical results which may provide an intuition for the discrepancies that we observe between our BLEU and ChRF computed scores and the ones presented in the literature of NMT. As loss value increases or drops close to 0, all the metric scores that we report do come from the first epoch of the training procedure. In that way our employed CMLM models do not get the chance to enchance their performance through iterative refinement and the sense of Auto - Regressive approach that provides.

Possible alternative reasons for these levels of translation quality could be: 1) learning rate warm up has been used which could lead to 0 rewards (hence will lead to 0 losses in many of the cases, since a reward is used for all the tokens in the same sentence). 2) The expected input of the reward metrics could be different than what is initially used as an input (which led to many 0 rewards). 3) Both imported metrics contain an ngram parameter (by default 4) which could be a possible explanation of the many 0 rewards. 4) The calculated rewards of for example the torchmetrics sentence-level BLEU and the NLTK sentence level BLEU were different, meaning that the calculation was not done in a correct manner. 5) The aggregation of the losses of all tokens at the end is not implemented correctly.

One possible solution to further encourage the performance of the model is a simultaneous use of various rewards in the reinforcement loss of a CMLM MRT model. For example sentence-level metrics such as BLEU, ChrF and TER could be calculated simultaneously (since these are string-based metrics) and thereafter aggregated to get a reward over three metrics by for exampling taking the mean of these rewards. On one hand this could possibly encourage the model's performance, while on the other hand this will most likely worsen the model's speed.

Regarding future work, an interesting research topic in the field of NMT would be to use at the same time a combination of a string based and a pretrained quality translation metric in the reinforcement learning reward part of the MRT loss function. Will this approach combine the benefits that both each family of metrics brings to the table or will it worsen models performance and just increase its latency?

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Marjan Ghazvininejad, Omer Levy, Yinhan Liu, and Luke Zettlemoyer. 2019. Mask-predict: Parallel decoding of conditional masked language models. *arXiv preprint arXiv:1904.09324*.

Junliang Guo, Xu Tan, Di He, Tao Qin, Linli Xu, and Tie-Yan Liu. 2019. Non-autoregressive neural machine translation with enhanced decoder input. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 3723–3730.

Tom Kocmi, Christian Federmann, Roman Grundkiewicz, Marcin Junczys-Dowmunt, Hitokazu Matsushita, and Arul Menezes. 2021. To ship or not to ship: An extensive evaluation of automatic metrics for machine translation. *arXiv preprint arXiv:2107.10821*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Chenze Shao, Yang Feng, Jinchao Zhang, Fandong Meng, and Jie Zhou. 2021. Sequence-level training for non-autoregressive neural machine translation. *Computational Linguistics*, 47(4):891–925.

Shiqi Shen, Yong Cheng, Zhongjun He, Wei He, Hua Wu, Maosong Sun, and Yang Liu. 2015. Minimum risk training for neural machine translation. *arXiv preprint arXiv:1512.02433*.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27.

Zhixing Tan, Shuo Wang, Zonghan Yang, Gang Chen, Xuancheng Huang, Maosong Sun, and Yang Liu. 2020. Neural machine translation: A review of methods, resources, and tools. *AI Open*, 1:5–21.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

John Wieting, Taylor Berg-Kirkpatrick, Kevin Gimpel, and Graham Neubig. 2019. Beyond BLEU:training neural machine translation with semantic similarity. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4344–4355, Florence, Italy. Association for Computational Linguistics.

Lijun Wu, Fei Tian, Tao Qin, Jianhuang Lai, and Tie-Yan Liu. 2018. A study of reinforcement learning for neural machine translation. *arXiv preprint arXiv:1808.08866*.