Appendix: Mixed Cross Entropy Loss for Neural Machine Translation

Haoran Li [*1] **Wei Lu** [*1]

Table 1. BLEU scores of Transformers trained with SKD on the WMT'16 Ro-En validations sets. The value in the bracket is the scale factor σ.

	CE	MIXED CE	SKD(0.001)	SKD(0.01)
BLEU	32.17	**32.72**	31.51	31.32
	SKD(0.1)	SKD(0.5)	SKD(1)	SKD(2)
BLEU	31.71	31.71	31.73	31.7

A. Comparison with Self-Knowledge Distillation (SKD)

Hahn & Choi (2019) proposed a method called self knowledge distillation similar to mixed CE in teacher forcing, where the coefficient α is computed from a scaled Euclidean distance between the embedding of the target token and model's greedy prediction. We re-implemented their method and tried different scale factors σ. The results are shown in Table 1. In our setting, SKD did not outperform the baseline (CE). We conjecture this is because the model to balance the weights of the gold token and the model's greedy prediction. However, in mixed CE, this is done by simply assigning α a linear increasing value.

B. Replacing mixed CE with Iterative training

Recall that in scheduled sampling, mixed CE consists of two parts and we optimize them simultaneously. Now we modified the training procedure: 1) in i-th iteration, we maximize the the first part in the first pass with the coefficient being 1; 2) in $i+1$-th iteration, we maximize the second part in the second pass with the coefficient being 1 as well. The results are shown in Table 2. Iterative training does not achieve comparable performance with mixed CE but is not worse than CE. We think the input change at adjacent iterations, i.e. gold input sequence at iteration i vs. mixed input sequence at iteration $i+1$, may be responsible for the lower performance. There might be some more stable training procedures and we leave this for future research.

C. Minimum Risk Training

We implemented Minimum Risk Training (Shen et al., 2016) following Edunov et al. (2018) with beam size 8. The results on the WMT'16 Ro-En are shown in Table. 3. We can see that MRT is better than CE but the MRT is significantly slower than CE and mixed CE due to the sampling procedure during training. It is relatively easier to finetune the hyperparameters of MRT on WMT'16 Ro-En, but it is prohibitive to do so on a larger data set due to the slow and unstable training.

Table 2. BLEU scores of Transformers trained with *iterative training* on the WMT'16 Ro-En validations sets.

	CE	MIXED CE	ITERATIVE
BLEU	32.66	**33.64**	32.8

Table 3. BLEU scores of Transformers trained with *minimum risk training* on the WMT'16 Ro-En validations and test sets.

	CE	MIXED CE	MRT
VALID BLEU	32.17	**32.72**	32.68
TEST BLEU	31.10	**32.17**	31.10

D. Domain Robustness

We also did some exploratory experiments of mixed CE + word oracle method in domain robustness problem, following Wang & Sennrich (2020). The models were trained on medical domain and tested on medical, IT, koran, law, subtitles domains. We list our preliminary results in Table 4. Note that our pre-processing steps are different from theirs and this results in the different results in CE training. We encourage the readers to explore more of mixed CE in domain robustness problem.

Table 4. BLEU scores of Transformers trained with *CE* and *mixed CE + word oracle* on the DE-EN OPUS data (Lison & Tiedemann, 2016). The "Dev" BLEU is given by the best single checkpoint while other "Test" BLEUs are given by the average model (average of the top-5 checkpoints).

	DEV (MEDICAL)	TEST (MEDICAL)
CE	60.94	56.32
MIXED CE+WO	**61.46**	**57.68**
	TEST (IT)	TEST (KORAN)
CE	11.50	0.8
MIXED CE+WO	**14.80**	**0.96**
	TEST (LAW)	TEST (SUBTITLES)
CE	14.74	2.13
MIXED CE+WO	**19.76**	**2.65**

References

Edunov, S., Ott, M., Auli, M., Grangier, D., and Ranzato, M. Classical structured prediction losses for sequence to sequence learning. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 355–364. Association for Computational Linguistics, 2018.

Hahn, S. and Choi, H. Self-knowledge distillation in natural language processing. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pp. 423–430, 2019.

Lison, P. and Tiedemann, J. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pp. 923–929, 2016.

Shen, S., Cheng, Y., He, Z., He, W., Wu, H., Sun, M., and Liu, Y. Minimum risk training for neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pp. 1683–1692, 2016.

Wang, C. and Sennrich, R. On exposure bias, hallucination and domain shift in neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 3544–3552, 2020.