神经机器翻译推断阶段信心校准研究

本论文由腾讯 AI Lab 和清华大学合作完成，作者提出了一种评估神经机器翻译模型在推断场景下信心校准偏差的方法，并发现 Transformer 模型的信心尽管在训练场景中校准较好，但是在推断场景中仍然存在较大的校准偏差。以下为论文的详细解读。

On the Inference Calibration of Neural Machine Translation

基于概率的机器学习模型在给出预测结果的同时，往往会输出一个对应的信心指数(i.e., confidence)，该信心指数可以代表模型对自身预测结果的正确性的一个估计。在金融、医疗等安全等级较高的场景中，我们希望模型不但有较好的预测精度(i.e., accuracy)，并且能够做到“知之为知之，不知为不知”，对预测结果的正确性有准确的估计。

我们可以设想一个场景：在一个共同抗击疫情的各国联合医疗队中，各国医护人员可以使用机器翻译系统进行交流。在涉及患者病情的关键性描述中，我们要求机器翻译系统要如实反映其对翻译结果的信心。对于模型不自信的翻译结果，我们可以请语言专家有针对性的进行后处理，对于大部分模型自信的结果，我们可以直接使用。由此可见，对自身输出结果是否有一个准确的信心估计，是衡量机器翻译模型能否实际部署的重要性质。

量化模型对预测结果信心校准偏差的前人工作大多是在分类任务上开展的。但是，不同于分类任务的单一输出，包括机器翻译在内的生成式自然语言任务的输

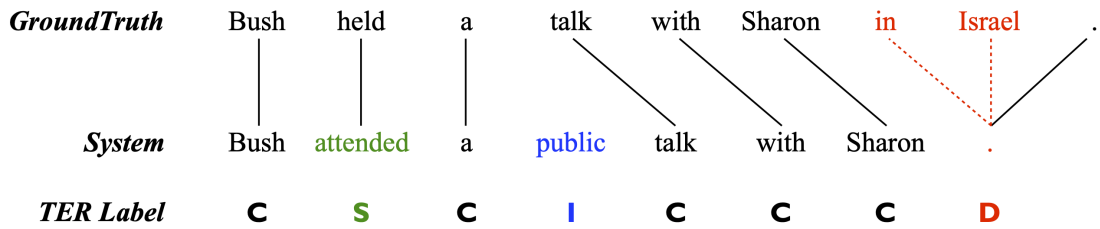
出都是序列化的，并且往往具有潜在的语义结构。如何评估序列化生成模型的信心校准偏差依然是一个尚未解决的问题。

在本文中，我们对期望校准偏差(Expected Calibration Error, ECE)进行了扩展，使其能够应用到序列化生成任务中来。具体地，ECE 在计算方式如下：

$$ECE = \sum_{m=1}^{M} \frac{|B_m|}{N} \left| acc(B_m) - conf(B_m) \right|$$

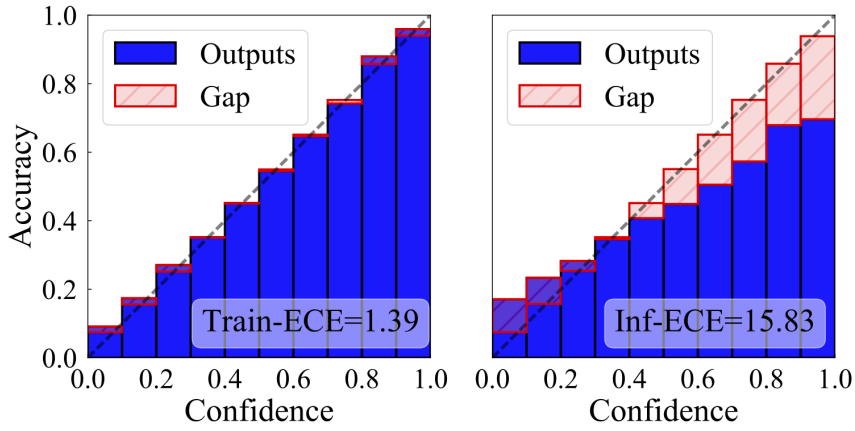
我们首先将模型在测试集中所有预测的 token 分为 M 组，分组的标准是每个 token 对应的信心指数（具体地，我们使用模型的翻译概率作为信心指数），信心指数相近的 token 会被分到同一组。在每一组中我们计算所有 token 的平均准确率和平均信心指数。对所有组的平均准确率与平均信心指数的偏差进行加权平均，将会得到最终的 ECE 结果。

为了计算 ECE，一个关键是如何量化每个 token 的准确性。为此，我们使用 TER 方法在模型译文和参考译文之间建立一个对应关系，并根据 TER 的标注决定每个 token 的正确性：



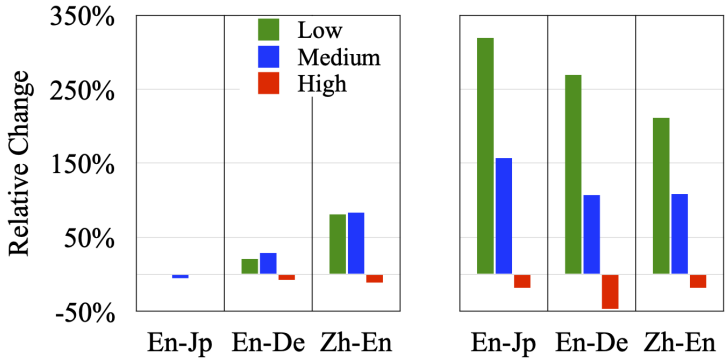
至此，我们就可以使用 ECE 量化序列化生成模型在推断场景下的信心校准偏差了。

在实验中，我们比较了机器翻译模型分别在训练与推断场景下信心校准偏差的情况：



(a) Training (b) Inference

可以看到模型在推断阶段的 ECE 远远高于在训练阶段的 ECE (15.83 > 1.39)，说明推断阶段的信心校准偏差对目前的机器翻译模型来说仍然是一个问题。为了深入理解模型信心校准的特性，我们分析了信心失准的 token 的语言学性质。首先，我们比较了不同频率的 token 的特性：



(a) Over-Estimation (b) Under-Estimation

实验发现模型在高频词上更不容易发生信心失准，而在中低频词上更容易发生信心失准。我们从相对位置、繁殖力、词性、词粒度等角度分析了模型的信心校准情况，详情请见论文。

为了探究当前深度学习技术与模型信心校准性质的影响，我们受 Guo et al., 2017 的启发，研究了正则化技术对机器翻译模型的影响：

Label Smoothing	Dropout	Beam Size = 10				Beam Size = 100			
		BLEU	*ECE*	*Over.*	*Under.*	*BLEU*	*ECE*	*Over.*	*Under.*
×	×	23.03	25.49	58.3%	9.6%	22.90	26.46	59.4%	9.3%
✓	×	24.51	14.99	42.3%	17.3%	24.58	15.97	42.8%	16.9%
×	✓	27.52	20.75	52.3%	10.1%	26.93	22.57	53.6%	9.8%
✓	✓	27.65	14.26	39.7%	14.1%	27.68	14.75	40.1%	14.2%
GRADUATED	✓	**27.76**	**5.07**	29.1%	31.6%	**28.07**	**5.23**	29.5%	31.4%

实验发现，dropout 和 label smoothing 这两个在 Transformer 模型中非常常用的正则化技术有利于降低模型的 ECE。基于实验发现，我们提出了一种 Graduated label smoothing 的方法，可以进一步减小模型在推断场景下的 ECE。具体地，我们的设计思想是对训练集中模型本身预测概率较高的样例使用较大的 smoothing 系数，对于预测概率较低的样例使用较小的 smoothing 系数。

我们还分析了 ECE 与模型大小的关系：

Enc.	Dec.	Para.	Beam Size = 10				Beam Size = 100			
			BLEU	*ECE*	*Over.*	*Under.*	*BLEU*	*ECE*	*Over.*	*Under.*
BASE	BASE	88M	27.65	14.26	39.7%	14.1%	27.68	14.75	40.1%	14.2%
DEEP	DEEP	220M	28.86	14.99	40.3%	14.1%	28.64	15.55	41.8%	14.0%
DEEP	BASE	145M	**29.09**	**14.28**	39.6%	14.1%	**29.29**	**14.53**	39.6%	14.2%
WIDE	WIDE	264M	28.66	16.09	42.3%	12.6%	28.42	17.22	43.2%	12.5%
WIDE	BASE	160M	**28.97**	**14.83**	39.7%	13.6%	**29.09**	**15.06**	39.8%	13.7%

实验发现尽管增大模型会提高翻译的 BLEU 值，但是也会导致模型的 ECE 升高，这是增大模型参数量的一个弊端。另外我们发现这个问题可以通过只增大编码器，保持解码器不变这一简单策略在一定程度上缓解。