IWSLT 2021语音翻译比赛总结

1. 比赛概况

国际上最具影响力的口语机器翻译评测比赛，至今已举办18届。[IWSLT 2021](https://iwslt.org/2021/)任务包括同传语音翻译、离线语音翻译、多语种语音翻译和低资源语音翻译四个赛道。

1. 任务简介

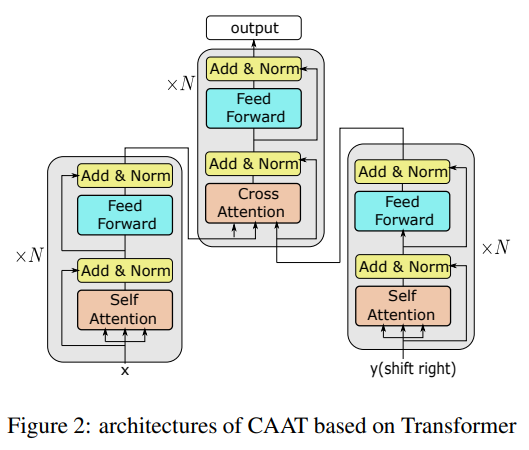
今年我们参加的是同传语音翻译赛道，同传赛道又包括三个子任务：英德文本翻译、英日文本翻译和英德语音翻译，本人负责英德语音翻译任务的ASR和pipeline系统，之后的总结都针对该子任务。

1. 数据集

英德ASR任务用到的原生数据集包括MuST-C-v2 (448h)、Europarl (155h)、CoVoST2 (1090h)、LibriSpeech (960h)和TED-LIUM3 (452h)，总计约3100h，经过data distill、SpecAugment并加入语音合成数据之后，用以ASR训练的全量数据集达到30000h。

1. 主要评测指标
2. BLEU（BiLingual Evaluation Understudy）：评价机器翻译质量的官方指标，越高代表翻译质量越好；
3. AL（Average Lagging）：平均延迟，定义为输入语音结束时刻与翻译结束时刻之间的时间差的均值，越低代表系统延迟越小，比赛官方将提交的语音翻译同传系统根据延迟划分低中高三档，标准分别为1000、2000和4000（ms）。
4. 模型结构和方法

所有三个子任务均使用新提出的Cross Attention Augmented Transducer (CAAT) 结构，如下图所示：



该模型结构基于Transformer，借鉴RNN-T结构实现online ASR。整体结构分为三部分：encoder、decoder和jointer，其中1）encoder为unidirectional Transformer encoder，主要思想为切block和利用fixed size right context和infinite left context做attention，具体实现包括：将input feature h切分为block，block step为m，block size为m+r (main + right)，对于第n个block，其main context为 ，right context为 ,于是attention的输入可以写成 , , ； 2）decoder结构类似Transformer decoder，运用masked self-attention保持其autoregressive的特性，不同点在于全为self-attention，原有的cross-attention被划到了jointer中； 3）jointer结构同Transformer decoder中对应cross-attention的部分，如若就此实现CAAT，该jointer中的计算复杂度将会远高于RNN-T，达到O(|x|·|y|)，故而定义了一个decision step size d，相当于在源端输入x上做了一次factor为d的下采样，从而将复杂度降到O(|x|·|y|/d)，类似RNN-T的输出维度（B T U V），CAAT jointer 的输出维度为（B G U V, G=T/d），除了降低复杂度外，定义此step size更重要的目的是为了控制系统延迟（step size d越大，context信息越丰富，性能越好，但延迟也就越高）。

训练使用multi-task training，loss组成包括CAAT\_loss、CTC\_loss和latency\_loss，其中CAAT\_loss计算方法同RNN-T, CTC\_loss由encoder引出，latency\_loss则是对CAAT解码路径（lattice path）中每一个代表non-blank输出的采样点做延迟惩罚，计算公式为：

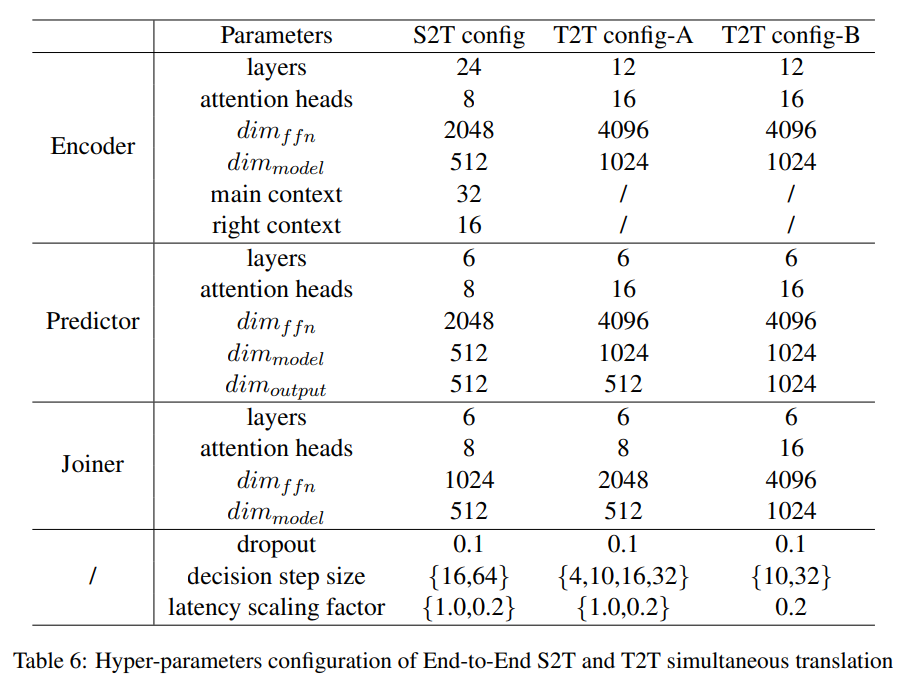
其中i，j分别代表该采样点对应x，y的下标，其归一化后的差距可以显式地作为当前延迟的衡量指标，然后对每条路径上的所有采样点做均值，最后根据forward-backward算法计算该指标在所有可能的解码路径上的期望，即为latency\_loss。

1. 延迟控制策略

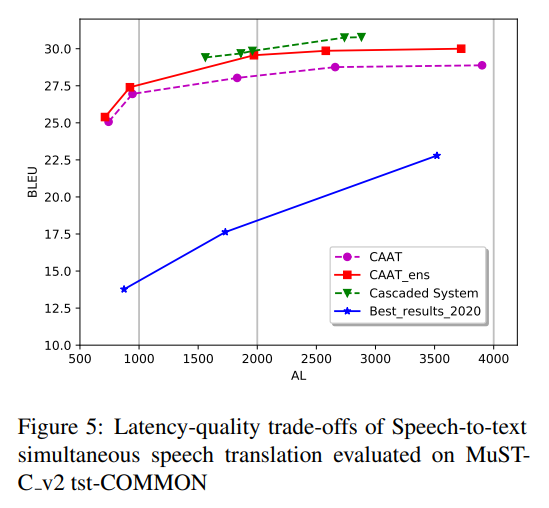
所有的三个子任务均使用simuleval工具进行测试，指标包括BLEU和AL，控制延迟的方法除了模型训练可调的超参step size（或downsample factor）外，在simuleval的解码逻辑中可通过控制每个step送入的语音帧数/文本token数（方便区分，这里称作test step size），来控制系统AL，以获得低中高每个档内的最优BLEU。以上所述适用于end2end和cascaded system，对于后者还可以通过控制级联时ASR与MT之间token flow来实现不同的系统延迟，具体实现有以下三种方法：1）greedy，每次ASR所有输出直接送入MT system，该策略会使得整个系统的延迟完全为ASR系统所主导，高延迟MT系统的优势发挥不了；2）step size，同样预定义一个超参d，ASR输出token累计数目到d才会被送入MT system，该策略可以充分利用高延迟MT system的优势，但部分系统延迟会超过上限4000ms；3）ranged step size，为2）中的step size定义一个范围 [lower, upper] ，当前累积token数落入范围内时会全部送出，如果低于下界会继续wait，高于上界时将token list平分然后依次送入MT system（此操作是为了避免送入过长的subsequence导致延迟爆炸和翻译质量显著下降，也可以采用其他的分段策略）。实验结果表明，方案3）会在所有的配置上获得最优性能（指对应延迟档的最高BLEU）。

1. 模型配置与实验结果

最优ASR模型配置与[paper](https://github.com/ustcychu/iwslt2021_simul_speech_trans/blob/main/iwslt_submission.pdf)中Table 6 的 “S2T config” 一致，其中 (decision step size, latency scaling factor) = (64, 0.2)，其他详见paper。



级联系统在中高延迟上的最优性能分别为：BLEU 29.85 / AL1959.03、BLEU 30.78 / AL 2881.20，比end2end对应的最优配置分别高0.29和0.78 BLEU；此外，该级联系统中的ASR系统性能为 WER 6.30 / AL 1720.20，MT系统性能为 BLEU 34.79 / AL 5.93。上述所有评测结果均基于MuST-C-v2 的tst-COMMON测试集。



如上图所示，对应同传语音翻译子任务，级联系统仍以一定的优势领先端到端系统，但与几年前级联系统一枝独秀的情况已大不一样。总体而言，我们的同传系统在性能上碾压了前两年最好的系统。

1. 比赛结果

待定

1. 总结

整个比赛过程中，主要负责基于CAAT的ASR system的搭建和调式，cascaded system的整体逻辑实现，以及最后阶段整个系统的打包调试任务等。