Week3

通常，我们认为 NLP = NLU + NLG，NLU(Neural Language Understanding)指的自然语言理解，NLG(Neural Language Generation)指的自然语言生成。NLU 负责理解内容，NLG 负责生成内容。

NLG 的主要目的是降低人类和机器之间的沟通鸿沟，将非语言格式的数据转换成人类可以理解的语言格式。

1.NLG 有 2 种方式：

text – to – text：文本到语言的生成

data – to – text ：数据到语言的生成

在现实应用中，可以把其应用在翻译、对话系统、文本摘要、数据报表的分析解读、图文文本描述(Image captioning)和写文章等领域。

NLG 任务大部分都是 seq2seq(Image captioning 除外)结构，decode 阶段为自然语言生成部分。其原理是把之前 encode 阶段产生的输出作为 decode 的输入，在每个时间步，生成对 vocab 里每个词汇的得分，然后通过 softmax 函数，得到每个 token 的概率，再通过相关函数，生成一句话。

2.Decode 主要算法：

Exhaustive Search Decoding：类似穷举法，对每个时间步生成的所有 token 概率进行追踪，再计算 T 个(句长)时间步生成的句子的概率，选择最优概率，但时间复杂度太贵；

Greedy Decoding(贪心算法)：选择每个时间步生成的 token 里概率最优的那个最终形成的句子，但此算法为局部最优解，非全局最优解；

Beam Decoding：此算法是基于 Greedy Decoding 的优化，选择每个时间步生成的 tokens 里的 top-k，k 为 beam size，再对选择的 tokens 生成的 hypothesis 获得得分。

最终可选择得分最优的。但此算法若选择的 k 太大，一则太贵，二则如在 NMT(翻译任务)中，倾向生成短句子，降低 BLEU 得分，三则生成的句子会越来越泛，和前文相关性变小。需合理选择 k 值；

在 decoding 任务中，会有倾向短句子和重复句子生成的问题。针对前一个问题，可通过归一化句子长度来解决此问题。针对重复句子生成问题，可人工选择非重复 n-grams，或者加入针对 h_t 相似性的惩罚项，再或者通过非似然估计法，惩罚已生成的 tokens，降低一些已生成 tokens 的概率。也可通过 F^2softmax 方法。

F^2softmax：此方法是先对数据集中出现的 tokens 进行频率计算，然后根据频率的高低，对 tokens 进行分组。在预测时，先预测 tokens 所在的频率组，再对每组里的 tokens 进行概率分布计算，选择 tokens；

Sampling Strategy(采样策略)：人工语言和 Beam Search 算法所产生的语言相比，会发现 Beam Search 存在随机性较低的问题，因此考虑在算法中引入一定的随机性。考虑对概率分布生成的 token 进行随机采样，但可能会采样到概率过低的词。为了避免太过随机，可通过 Temperature Scaling 参数，放大随机概率大的 token 概率，同时，减小随机概率小的 token 概率。也可通过 Top-k sampling 方法。

Top-k sampling：可通过 tokens 概率分布的概率高低，选择 top-k 个 tokens。但此方法会因为概率分布本身的原因，如扁平的概率分布可能会导致放弃一些概率还可以的 tokens 或者峰值分布会选择到一些概率很低的 tokens。为避免此问题，可采用 Top-p(Nucleus) Sampling 方法，按概率分布的百分比对 tokens 进行选择。

3.度量指标：

主要的度量指标有 ROUGE-N(统计 n-gram)和 ROUGE-L(统计最长子串)。

参考文献：

1. 《自然语言生成-Natural-language generation》，产品经理的 AI 知识库，https://easyai.tech/ai-definition/nlg/
2. 其他内容参考课件