|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 论文题目 | Understanding the Properties of Minimum Bayes Risk Decoding in Neural Machine Translation | | | |
| Paper URL | https://aclanthology.org/2021.acl-long.22/ | | | |
| Project URL | https://github.com/ZurichNLP/understanding-mbr | | | |
| 综述/背景介绍 | 发展状况 | 原因 | 意义 | 关键词（速记词汇、信息索引词汇） |
| 神经机器翻译(NMT)目前存在着一些问题，例如低估了翻译的真实长度，低估了罕见单词和过度生成非常频繁的单词的概率，或者容易受到训练数据中的复制噪声的影响。在域外翻译中，幻觉(流利但与来源无关的翻译)很常见。以前的工作已经用诸如长度标准化(Wu等人，2016年)、数据清洗或模型正则化。 | | |  |
| 假设 | Eikema和Aziz(2020)强调了决策规则的作用，即寻找得分最高的翻译，并认为它至少要为这些偏见和缺陷承担部分责任。他们发现，来自NMT模型的采样与训练数据统计数据是一致的，而beam search则不是。他们建议该领域研究基于无偏样本的替代推理算法，例如最小贝叶斯风险(MBR)解码。  我们相信MBR有潜力克服NMT的几个已知偏见。更准确地说，如果偏差可以被理解为是由beam search的模式搜索性质引起的，那么我们假设MBR可能表现出较少的偏差。我们认为短小的翻译、原文的副本和幻觉是可能的假设，但与其他可能的假设有很大的不同。如果这样的病理性假设是在一个样本池中，MBR不太可能选择它们作为最终的翻译。 | | |  |
| 方法描述(含图) | 虽然Eikema和Aziz(2020)比较了样本和beam search输出的统计特性，并表明根据自动度量，MBR可以比beam search性能更好，但我们的目标是对MBR及其特性进行有针对性的研究，特别是它对前面讨论的偏差和缺点的影响。 | | |  |
| 实验设计 |  | | |  |
| 数据处理 | 输入 | 处理方式 | | 关键词（速记词汇、信息索引词汇） |
| 语句 | 我们使用Sentencepiece (Kudo, 2018)，子词正则化作为唯一的预处理步骤，它负责标记化和子词分割。词汇表中所需的片段数量随数据集的大小而变化。 | |  |
| 结论 | 在我们的实验中我们发现：  如果与有利于短翻译的效用函数一起使用，则MBR继承了这种偏差；MBR仍然表现出token概率偏差，因为它低估了稀有token的概率，而高估了非常常见的token；与beam search相比，MBR解码对训练数据中的复制噪声更稳健；与beam search相比，MBR表现出更高的域稳健性。我们证明，MBR减少了翻译中幻觉内容的数量。 | | | |
| 局限性分析 |  | | | |

**（论文名）：**理解神经机器翻译中最小贝叶斯风险译码的性质

**（题目）：**《Understanding the Properties of Minimum Bayes Risk Decoding in Neural Machine Translation》

**（论文URL）：**论文地址: https://aclanthology.org/2021.acl-long.22/

**（总结）：**MBR译码作为一种决策规则，由于具有克服NMT中MAP译码的某些偏差的潜力，最近在MT中重新受到关注。我们实证研究了以常用MT度量为效用函数的MBR译码特性，发现它仍然表现出与beam search相似的长度偏差和token频率偏差。长度偏差与效用函数密切相关。然而，我们也观察到，MBR解码成功地缓解了许多众所周知的NMT失败模式，如伪复制，或在域转移下的幻觉。MBR达到这种稳当性的机制是，样本池中的副本或幻觉假设被赋予较低的效用，从未被选择作为最终翻译。

在我们的实验中，根据自动度量标准，MBR并没有普遍优于beam search，但由于其鲁棒性，我们仍然认为它是一个有希望的替代MAP解码。在未来的工作中，我们有兴趣探索更复杂的相似度指标，以用作效用函数，包括可训练的指标，如COMET (Rei等人，2020)，并调查这些效用函数如何影响翻译的整体质量和偏差。

**（附图）：**

NMT中通常采用标准解码算法是beam search。beam search属于一类更广泛的推理过程，称为最大后验概率(MAP)算法。MAP算法的共同点是，它们都试图在给定模型下找到最可能的翻译。本质上，它们试图恢复序列上的输出分布模式。beam search是一种易于处理的近似值，但它也经常无法找到分布的真实模式。

长度偏差：系统低估了翻译的真实长度。平均而言，他们的译文比参考文献短

词频偏差：在翻译中，训练数据中频繁出现的标记被过度表示。另一方面，罕见的token出现的次数比训练数据中显示的概率要少

Beam search诅咒：增加光束大小会导致找到在模型下更有可能发生的翻译。从理论上讲，这应该会提高翻译质量。矛盾的是，经验结果表明，size过大会降低质量

对复制噪声的敏感性：训练数据中复制的内容不成比例地影响翻译质量。更具体地说，最有害的一种是训练数据目标侧的源句副本(Khayrallah和Koehn，2018年)。如果训练数据中存在这样的副本，则复制假设在Beam search中将被过度表示(Ott等人，2018年)。

域健壮性低：系统在分布移位(例如域移位)下不健壮。在未知的测试领域进行系统翻译通常不会逐渐降低翻译质量，而是会导致称为幻觉的完全失败案例(Lee等人，2018；Koehn和Knowles，2017；Müler等人，2020)。

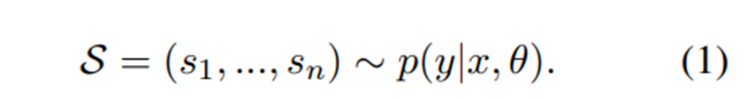
过去的许多研究将这些不足归因于模型架构或训练算法，而将beam search视为实验中的固定常数。相比之下，Eikema和Aziz(2020)认为模型的适合性是合理的，这意味着无论是模型本身还是它的训练都不会有错。相反，他们认为根本问题在于beam search。

模式的不足：Stahlberg和Byrne(2019)和Eikema和Aziz(2020)表明，输出序列上的分配模式实际上并不是最好的翻译。相反，在许多情况下，模式似乎是空序列(Stahlberg和Byrne，2019)。此外，该模式的概率似乎与许多其他序列没有太大区别，因为在输出空间的大范围内，输出分布相当平坦(Eikema和Aziz，2020)。直观地说，在NMT训练中可能出现这样的情况是有意义的：最大似然估计训练不会约束一个模型仅仅通过它的模式来很好地表征。如果模式不充分，那么显然这对于诸如beam search和地图推理之类的模式搜索过程是有问题的。事实上，只有当输出分布的模式是可信的时，才应该使用MAP解码(Smith，2011)。另一种选择是一种决策规则，它考虑翻译与其他可能的翻译的不同之处。

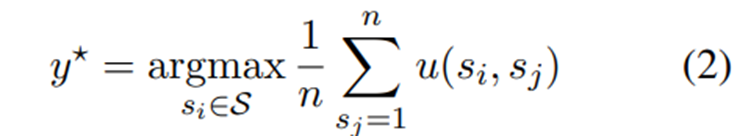
**最小贝叶斯风险解码**

MBR解码用于语音识别(Goel和Byrne，2000)和统计机器翻译(Kumar和Byrne，2004；Tromble等人，2008)。最近，MBR也被用于改进NMT中的beam search解码(Stahlberg等人，2017；Shu和Nakayama，2017；Blain等人，2017)。Eikema和Aziz(2020)是第一个测试MBR变体的人，这种变体对样本进行操作，而不是通过Beam search生成n 个Best列表。

从本质上讲，MBR的目标不是寻找最可能的翻译，而是对于给定的损失函数和真实的后验分布，找到最小化预期风险的翻译。实际上，所有可能的候选翻译的集合可以通过从模型中提取大小为n的样本池S来近似：

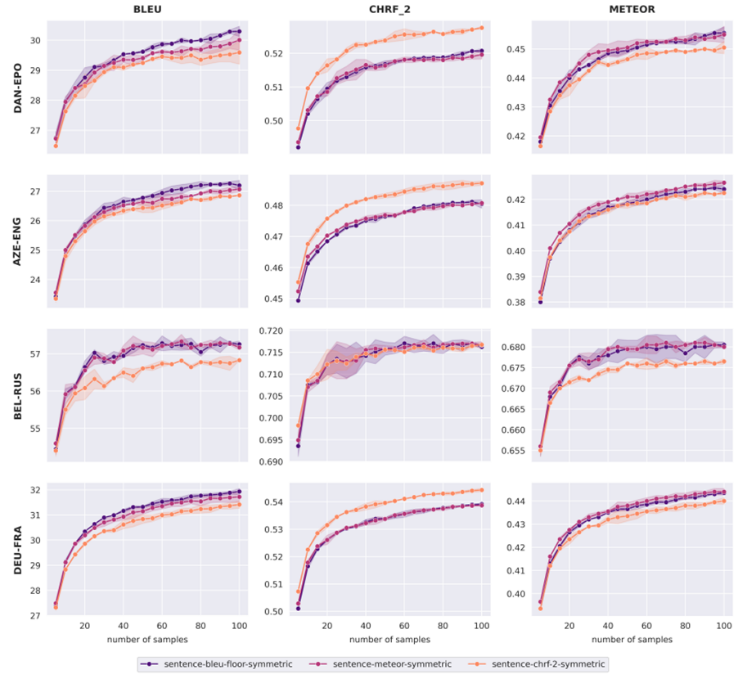


同样的一组样本也可以用来近似真实的后验分布。然后，对于S中的每个样本si，通过将其与池中的所有其他样本进行比较来计算其预期效用(逆风险)。选择预期效用最高的样本作为最终翻译：



1、长度偏差

我们用不同的实用函数来评估MBR解码。没有一个单一的效用函数在所有的评估指标上表现最好。相反，我们的任何评估指标都可以通过选择一个密切相关的效用函数来优化我们发现，随着样本数量的增加，MBR的翻译质量稳步提高。MBR不会受到Beam search的困扰，在大的搜索size中，单一的病理假设会危及翻译质量。



我们在下表中分析了不同解码方法产生的翻译长度。我们发现，在平均翻译长度方面，beam search低估了真实的翻译长度，即使假设是标准化的。抽样产生的假设与参考长度更匹配。这与Eikema和Aziz(2020)的研究结果一致。对于MBR解码，很明显，效用函数的选择对最终翻译的平均长度有影响。例如，使用句子级别的BLEU作为效用函数会导致翻译太短。BLEU是一种基于精度的衡量标准，在句子层面上更倾向于较短的翻译(Nakov等人，2012)。chrf-2和meteor更强调回忆，由此产生的MBR翻译高估了翻译的真实长度另一方面，CHRF -0.5是CHRF的一个变体，具有准确性的偏见，导致了最短的翻译。

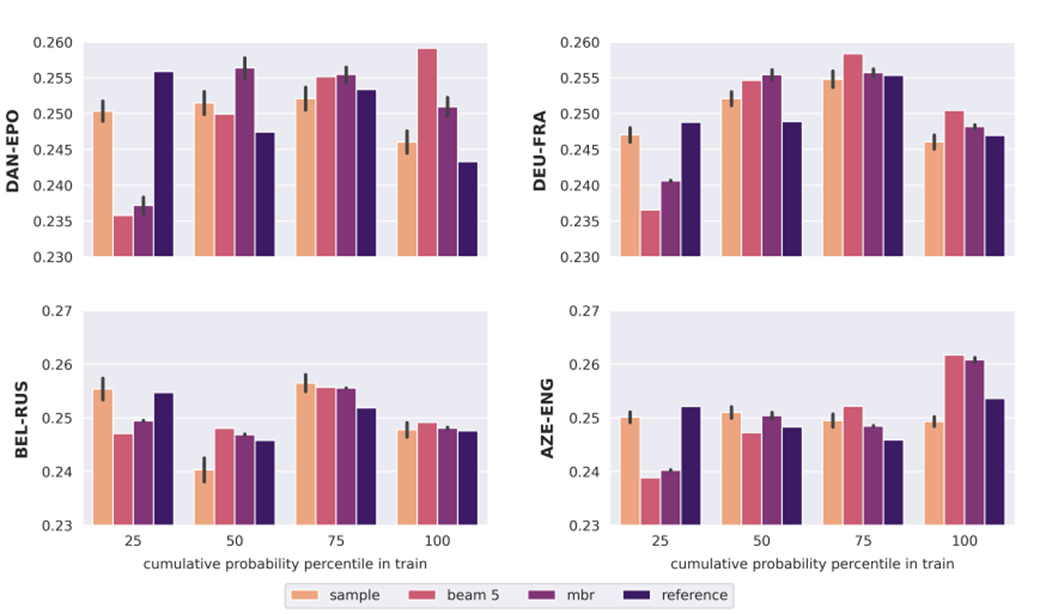
表格

描述已自动生成

**基于这些观察，我们得出结论，MBR继承了与其效用函数相关的长度偏差**

2、Token频率偏差

beam search过度生成在训练数据中非常常见的token，而不足生成稀有的token。另一方面，采样将正确的概率分配给常见和罕见的token。鉴于MBR是基于样本的，它是否与样本共享这一属性？



在图中，我们表明情况并非如此。尽管MBR的概率偏斜度没有beam search那么严重，但MBR仍然为频繁事件分配了过高的概率。这样做的一个原因是，我们的效用函数是基于样本之间的表面相似性的，因此很少采样的稀有标记也将具有较低的效用。不幸的是，非常常见和非常罕见的单词的正确概率统计与翻译质量之间存在权衡。抽样可以得到最真实的统计数据，但抽样会导致整体翻译质量最差。

3、领域鲁棒性

一般来说，随着样本数量的增加，MBR接近，但在我们的域内数据上的性能并不优于Beam search。下图显示在该基准测试上，MBR在4个未知测试域中的2个上的性能优于Beam search。MBR之所以能够在未知领域胜过Beam search，一个可能的原因是它减少了幻觉翻译。为了检验这一假设，我们将幻觉定义为在Lee等人的启发下，与参考文献相比，CHRF2得分低于0.01的翻译。

图表

描述已自动生成

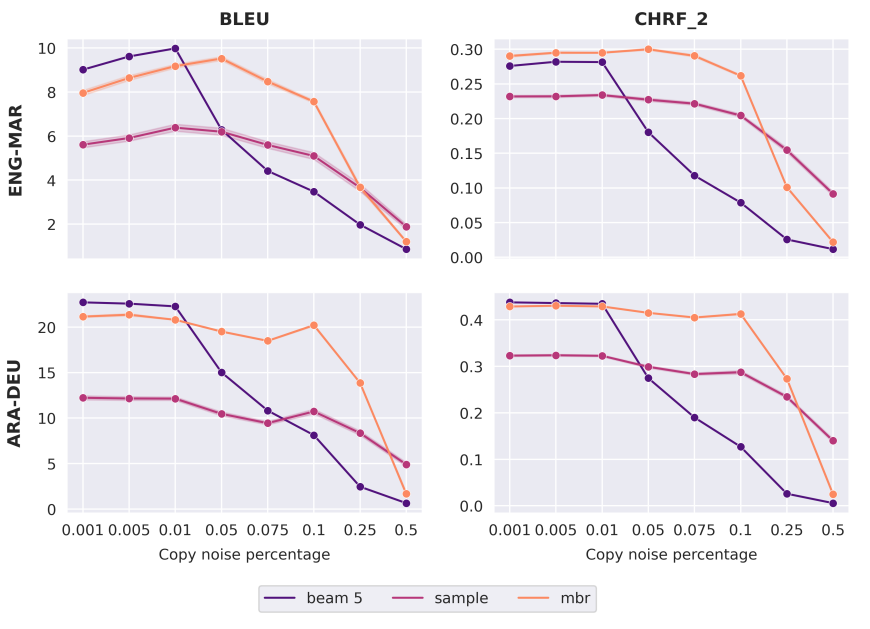
下图显示， MBR为幻觉的假设分配了较低的效用分数。同样，与光束搜索或抽样相比，MBR降低了在最终翻译中发现的幻觉百分比。综上所述，我们发现MBR译码比beam search具有更高的域稳健性。

图表, 条形图

描述已自动生成

4、复制噪声对训练数据的影响

如果在训练数据的目标侧存在源句子的副本，则在beam search中副本的比例过高。在这里，我们测试MBR是否也受到这种复制偏见的影响。如图所示，如果训练数据中的副本很少，则MBR和BEAM搜索具有可比性。然而，如果所有训练样本中有5%到25%是副本，则MBR的性能远远超过BEAM搜索(对于阿拉伯语-德语，BLEU>10)。



为了进一步证明MBR容忍复制噪声的能力，我们对副本进行了分析。我们将副本定义为引用大于0.9的单词重叠的翻译。我们表明，MBR分配给复制假设的效用比所有假设加在一起的效用要低得多。在最终的翻译中，MBR设法大幅减少了副本。例如，如果大约10%的训练样本是副本，BEAM搜索产生大约50%的副本，而MBR将这个数字减少到10%以下。

图表, 瀑布图

描述已自动生成

MBR对训练数据中的复制噪声具有更强的鲁棒性。我们承认这个设置是人为的，因为复制噪声可以很容易地从数据集中删除。然而，这是一个明显的例子，说明了NMT系统的一个已知缺陷，通常被归因于模型或训练程序，而事实上beam search至少是部分原因。