7.4 机器翻译

在本节最后，我们来看机器翻译问题的评估方法。类似的自然语言处理相关问题，比如问答系统、对话系统、看图说话问题等都需要一个评判结果的指标。最简单有效的方法当然是通过人工评估打分，但是数据量大起来以后每个模型都用人评不太现实，所以需要一个度量模型好坏的方法。大体可以通过两种方法评价，一种基于word，就是看目标和预测值有几个word重合；一种是基于word-embedding，即计算words之间embedding的相似度。基于word的方法BLEU (Papineni et al., 2002) , METEOR(Banerjee & Lavie, 2005) 和ROUGE (Lin, 2004)通常用于机器翻译。

基于word的方法：

预测 A: Mary is going to fly from B to A.

预测 B: Mary is going from A to B.

目标句: Mary is going to fly from A to B.

1. Precision & Recall

最容易想到的做法：

Precicion = num(重合词数) / length(预测句)

Recall = num(重合词数) / length(目标句)

其中，重合词数是指预测句和目标句的重合词。

评价：不靠谱。

预测A: precision = 1, recall = 1

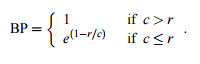
预测B: precision = 1, recall = 7/9

所以A比B翻译的好？

1. BLEU

IBM于2002年提出。

1. 用n-gram precision代替word precision. 最后的分数是所有n-gram precision的几何均值
2. 引入brevity penalty抑制太短的句子得分过高。由于n-gram precision天然对长句有惩罚作用（分母大），就容易产生短句得分过高的情况，为此

最终，BP=

BLEU =

评价：比较靠谱

以N-gram中最大N=3为例，

|  |  |  |
| --- | --- | --- |
| Metric | 预测A | 预测B |
| 1-gram precision | 9/9 | 7/7 |
| 2-gram precision | 5/8 | 5/6 |
| 3-gram precision | 4/7 | 3/5 |
| Brevity penalty | 9/9 | 7/9 |
| BLEU | 36% | 39% |

1. METEOR

METEOR全名：Metric for Evaluation of Translation with Explicit Ordering

1. ROUGE

**需要注意的是**，对话系统和机器翻译任务看似相似，可以同样采用这些方法，但效果并不好，因为对话系统需要判别给定对话中的有效回复数。详见<How **NOT** To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation>。该文指出在对话系统中（不管是retrieval-based 方法还是generation-based方法），尽管上述指标的值最近一直在涨，但它们和人工评价相关性并不强， 也就是不能真正度量对话系统的好坏。