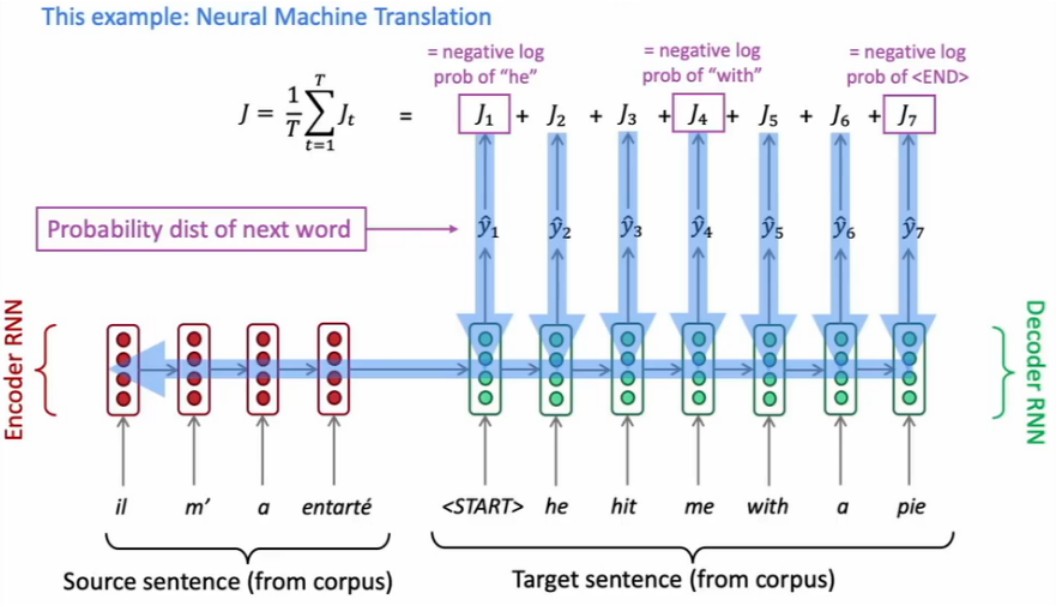
Lecture15 Natural Language Generation

创建时间：	2019/12/31 16:41	**更新时间：**	2020/1/13 21:03
作者：	wjj4work@163.com		

RECAP

- Language Modeling：预测下一单词
- Conditional Language Modeling：给定之前的单词和一些其他输入来预测下一单词
- 训练（conditional）RNN-LM：



  将gold目标序列输入解码器。黄金输入forceLM而不是在每一步使用自己的预测
- 解码算法：
  - greedy decoding：输出的结果很差，可能不合语法，或者不自然
  - beam search decoding：旨在找到高概率序列的搜索算法。在解码的每一步，追踪k个（beam size）最可能的部分序列（称之为hypotheses）；到达停止标准时，选择最高概率的序列作为输出。通常会提供比贪婪搜索更好的质量。
    过大增加k实际上会降低BLEU分数（大k会得到较短的翻译）。在对话任务中，大k会输出很通用的句子，但是相关性很小。
  - sampling-based decoding：
    - pure sampling：从概率分布中随机取样
    - top-n sampling：前n各最可能的词采样。n=1：贪婪搜索。
      增加n以获得更多样化/风险的输出；减少n以获得更通用/安全的输出
- Softmax temperature。降低会使概率分布变得更加尖锐；提高会使变得均匀

NLG

NLG是很多任务的子任务，如机器翻译，摘要，对话，写作，图像字幕，自由问答等

摘要

给定输入文本x，写出更短的摘要 y 并包含 x 的主要信息
摘要可以是单文档，也可以是多文档。
数据集：

- Gigaword: 新闻文章的前一两句→ 标题(即句子压缩)
- LCSTS (中文微博)：段落→句子摘要
- NYT, CNN/DailyMail: 新闻文章→ (多个)句子摘要
- Wikihow (new!): 完整的how-to文章→ 摘要句子

句子简化：用更简单的方法重写原文本
数据集：

- Simple Wikipedia：标准维基百科句子→ 简单版本
- Newsela：新闻文章→为儿童写的版本

有两种主要策略：提取摘要和抽象概括

- 提取摘要。选择部分原始文本形成摘要。更简单但严格
- 抽象概括。使用NLG技术生成一些新文本。非常困难

Pre-neural summarization

大多是抽取式的

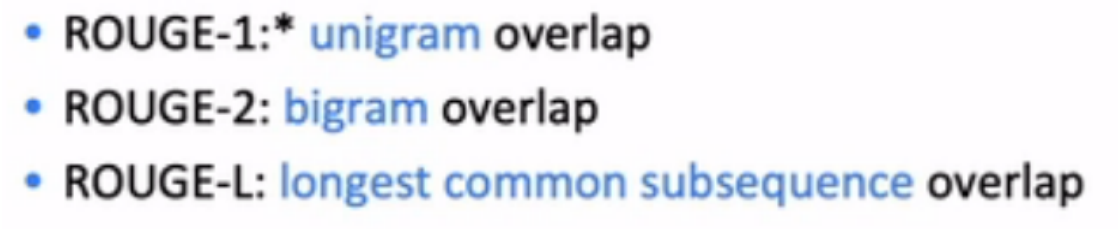
1. Content selection：选择一些句子
2. Information ordering：为选择的句子排序
3. Sentence realization：编辑并输出句子序列例如，简化、删除部分、修复连续性问题)

句子得分函数依据重要的句子，关键词，句子在段落中的位置等
图算法将文档视为一组句子(节点)，每对句子之间存在边，边的权重与句子相似度成正比。
使用图算法来识别图中最重要的句子。

GOUGE：摘要评估
Recall-Oriented Understudy for Gisting Evaluation
基于n-gram重叠。与BLEU的主要区别在于

- ROUGE没有简短的惩罚
- ROUGE是基于召回率（对摘要任务更重要），而BLEU基于准确度（对MT较重要）
- 通常使用F1值

可以用python实现



Neural summarization

2015: Rush et al. publish the first seq2seq summarization paper提出将抽象摘要看做翻译任务
2015年以后有了更多的发展：
copy mechanisms 基础的seq2seq+attention擅长流利的输出，但很难正确的复制诸如罕见词之类的细节。将注意力分配到想要复制的内容上。
问题：

- 复制太多，会崩溃为一个主要是抽取的系统
- 不善于整体内容的选择，特别是如果输入文档很长的情况下
  better content selection：
  pre-neural摘要是不同阶段的内容选择和表面实现而标准seq2seq + attention的摘要系统，这两个阶段是混合在一起的——解决办法：bottom-up summarization
  内容选择阶段：使用一个神经序列标注模型来将单词标注为 include / dont-include
  注意力阶段：seq2seq+attention系统不能处理 dont-include的单词（使用 mask ）

Reinforcement Learning：
A Deep Reinforced Model for Abstractive Summarization, Paulus et al, 2017
使用 RL 直接优化 ROUGE-L，获得了更高的ROUGE分数，但人类判断分数较低

Dialogue

对话的分类：

- 面向任务的对话
- 社会对话

pre-neural对话系统经常使用预定义的模板，或从语料库中检索一个适当的回答
Seq2seq-based dialogue：有严重的缺陷：

- 回应通用或无聊
  solution：可以直接在beam search过程中增加稀有词的权重；或使用sampling decoding；或使用检索和优化模型而不是从头开始生成模型。可以从采样中获得所有细粒度细节，然后跟警察需要编辑它以适应当前的情况
- 不相关的回应
  solution：改变训练目标。不是试图优化从输入S到相应T的映射使得最大化给定S的T

的概率，而是最大化最大相互信息

$$\log \frac{p(S,T)}{p(S)p(T)}$$

$$\hat{T} = \arg\max_{T} \left\{ \log p(T|S) - \log p(T) \right\}$$

- 重复
  solution：直接在 Beam 搜索中禁止重复n-grams。更复杂的方法是训练覆盖机制，可以防止注意力机制多次注意相同的单词；或是顶一个亿训练目标来阻止重复。
- 缺乏上下文，不记得对话历史
- 缺乏一致的角色

Storytelling

NLG evaluation

- 基于单词重叠的指标，例如BLEU、ROUGE、METEOR等不适合机器翻译，用于评估摘要任务、对话任务也很糟糕。
- 而困惑度能抓住语言模型有多强大，但不会告诉你关于生成的任何事情。
- 基于词嵌入的指标：更灵活，但与人类判断不相关。
- 可以定义更多的集中自动度量来捕捉生成文本的特定方面
  - 流利性(使用训练好的LM计算概率)
  - 正确的风格(使用目标语料库上训练好的LM的概率)
  - 多样性(罕见的用词，n-grams 的独特性)
  - 相关输入(语义相似性度量)
  - 简单的长度和重复
  - 特定于任务的指标，如摘要的压缩率
    虽然这些不衡量整体质量，他们可以帮助我们跟踪一些我们关心的重要品质
- 人类评估——黄金标准
  但有不一致、可能不合逻辑、注意力不集中、曲解问题的问题

current trends

- 将离散潜在变量纳入NLG
- 严格的从左到右生成的替代方案
- 替代teacher forcing的最大可能性训练。更全面的句子级别的目标函数（而不是单词级别）