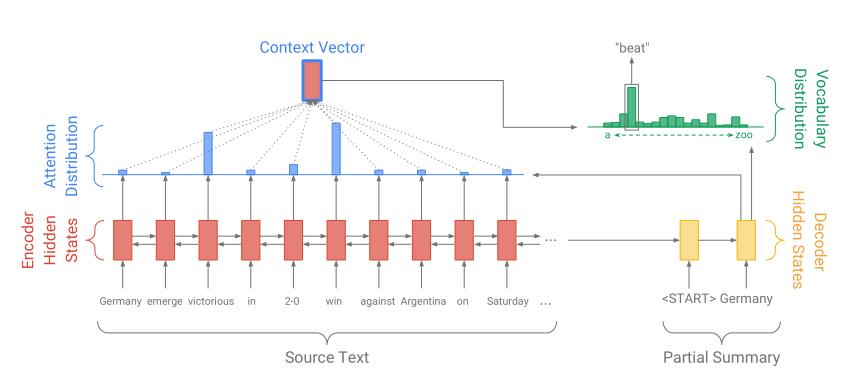
# 一、深度学习

## 1.1 NLP

### 1.1.1 摘要模型



Baseline sequence-to-sequence model with attention

基线：encoder侧的隐层h作为k、v，然后decoder侧的隐层s作为q，基于weight sum求出语境向量context vector，将context与st concate之后送入全连接层，求出预测的单词。

#### 1.1.1.1 文献阅读

*1、Get To The Point: Summarization with Pointer-Generator Networks-Stanford University-2017-指针网络*

**论文摘要：**

当前sequence2sequence modle 存在的问题与解决办法：

问题1:复制原文不准确的信息

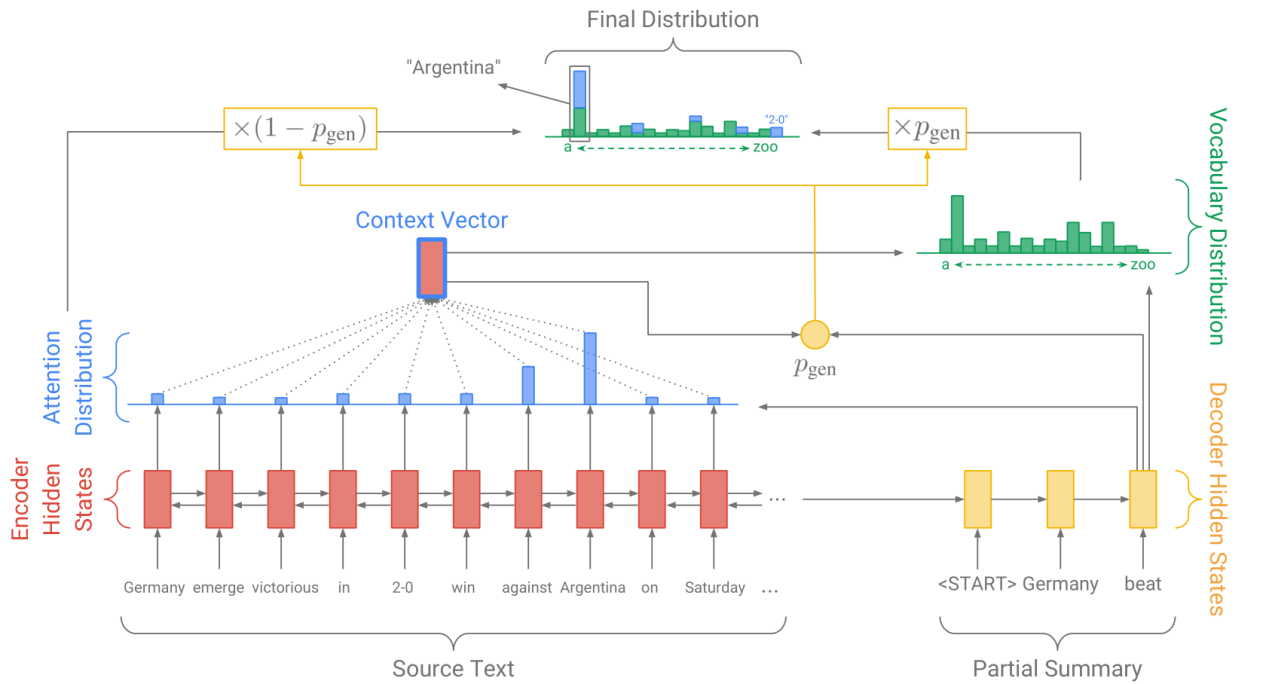
问题2:生成词自己重复

解决方案：pointer-generator network

1、pointer net：从原文复制使得能够准确的复现原文关键信息；另外generate net使得网络具有生成新词的能力；

2、利用coverage追踪已经生成的词语来缓解摘要重复问题。(从机器翻译引入)

**模型：Pointer-Generator-Model**



Pointer-generator model

attention参数和语境向量的计算和baseline一样，开关函数的公式为：



其中ht\*表示语境向量，st表示解码器输出，xt表示解码器输入，其他参数可以学习。每一个词语的生成概率公式为：



Pvocab表示全连接层输出的词表中每一个词的概率，ai为每个单词的attention得分。所以未登录词也会有得分。

覆盖机制：为了缓解解码时候的自我复制问题，借鉴MT(机器翻译)模型中的coverage机制，具体而言，在每一个解码时间步，维护一个覆盖权重：



attention机制也修改为：



其中h仍然是step的隐层向量，st为解码侧隐层向量，ct为覆盖向量。最后的分布权重仍然为at=softmax(et)。

损失函数修改为：



ct与at分别为历史总和与当前值，作为损失函数那么就希望覆盖向量变小，但是在每一个时间步at的和为softmax求和为定值1，所以对已经有很多注意力的位置不利。

**数据集**

CNN/Daily Mail Dataset：新闻内容平均781个词语，摘要平均56个词语，

287,226(28万)training pairs, 13,368(1万) validation pairs and 11,490(1万)test pairs

**实验**

1、并不是一开始就加上覆盖机制。一开始就假如覆盖机制的话会降低模型的表现。

2、只把覆盖向量添加到attention机制中，而不添加到损失函数中的话作用不大。

3、使用128维度的词向量，但是不是从头开始训练的，对比模型是从头开始训练的。

数据集：https://github.com/abisee/cnn-dailymail

python2代码：https://github.com/abisee/pointer-generator

python3代码：https://github.com/becxer/pointer-generator/

2、《基于改进Sequence-to-Sequence模型的文本摘要生成方法》-北京林业大学.周健.2019

**背景：**

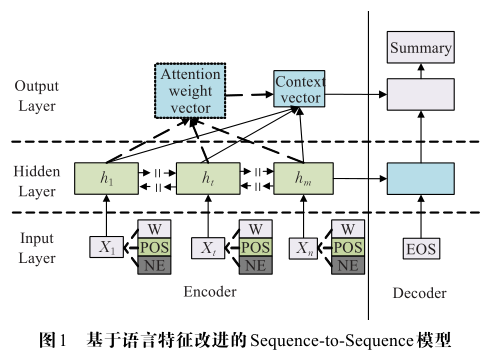
问题1：模型不能充分利用文本的语言特征信息； 对策1：加入语言特征

问题2：生成未登录词，影响阅读；对策2：利用copyNet缓解该问题

**什么是语言特征？**

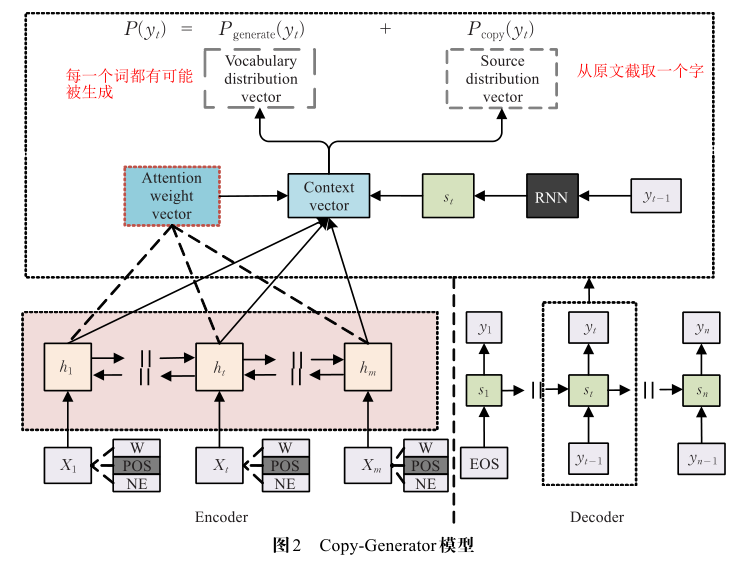
语言特征包括：词性、命名实体等，PS文本还具有统计特征包括tfidf、语义特征等等

**基本模型结构：**



Encoder和Decoder均采用RNN实现。首先在编码阶段，将词向量、词性类别向量、NE类别向量concate、然后就出每一个step的隐层向量。解码的时候，基本的做法是将encoder最后一个step作为隐层输入，然后逐个的生成摘要内容。引入attention的做法是：在每一个解码时间步，将解码隐层作为q，每一个编码隐层作为k、v，然后将两个向量concate过全连接求单词。

**copy-generator模型：**

****

在解码阶段，每个时间步的预测词yt可由两种方式产生：生成模式和拷贝模式。

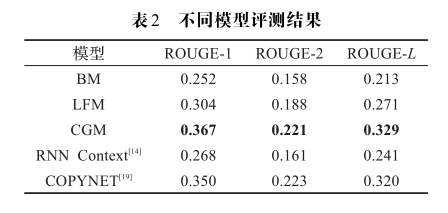
生成模式：基于encoder的隐藏状态和当前decoder步的隐层状态来求出预测词。

拷贝模式：利用两侧的注意力机制求内积，分数高的词被选中。

利用开关机制来决定模式选择：开关量的取值由编码隐向量、解码意向量

实验结果如何？和谁比较有提升？

数据集采用LCSTS，实验结果如下：其中BM表示编码和解码+attention模型，LFM表示BM+语言特征；CGM表示copy-generate model，CopyNet为经典模型。



词向量表示方法？

word2vec向量，非关键词用特殊标记代替：（1）NAME\_EN — 英文标记；

（2）NUM — 数字标记；（3）DATE — 日期标志；（4）UNK — 未登录词；

（5）EOS — 文本结束标记