心得2

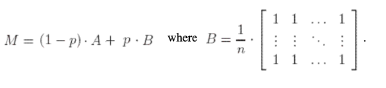
摘要分为生成式摘要和抽取式摘要。

抽取式摘要主要是将文章中一些重要的句子提取出去。它的其中一个重要算法是从Page Rank演变来的Text Rank。

Page Rank的主要思想是从网站的搜索权重来的。网站链接中互相引用从而体现网站的重要性。将这些互相链接转换成矩阵。矩阵中数值主要是一个网站中有多少指向其他网站的链接，然后将数值1平均分成几份，矩阵中Aij代表第i个网址出度到第j个网址。然后初始化一个V向量代表每一个网址的初始权重，然后重复进行AV,A(AV),A(A(AV)),直到这个乘积数值趋于稳定，稳定后的矩阵数值则代表每一个网址的权重。

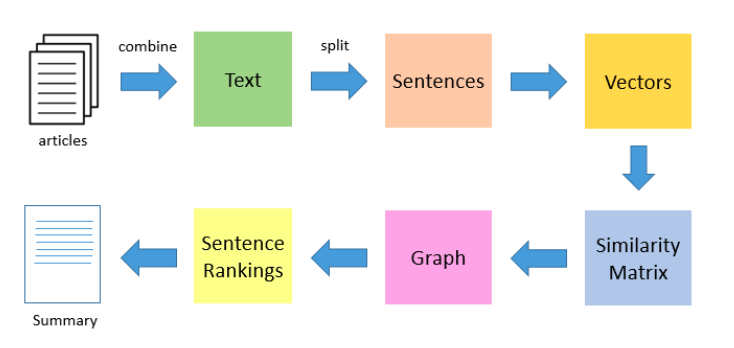
这个算法的缺点在于有一些网址的出度为0，则进行最后乘积算法的数值会变成一个零矩阵，这显然是不合理的，因为它可以有入度，也有重要性。解决方案在于我们可以设置一个相关的节点的向量全部为1/3。

二，有些网址是两两互不相连的，这就造成了权重矩阵是一个分块矩阵，最后乘积稳定变成一个拥有多个0的向量，这也显然是不合理的，我们可以修补一个在0到1的常量p,page的矩阵就变成下面

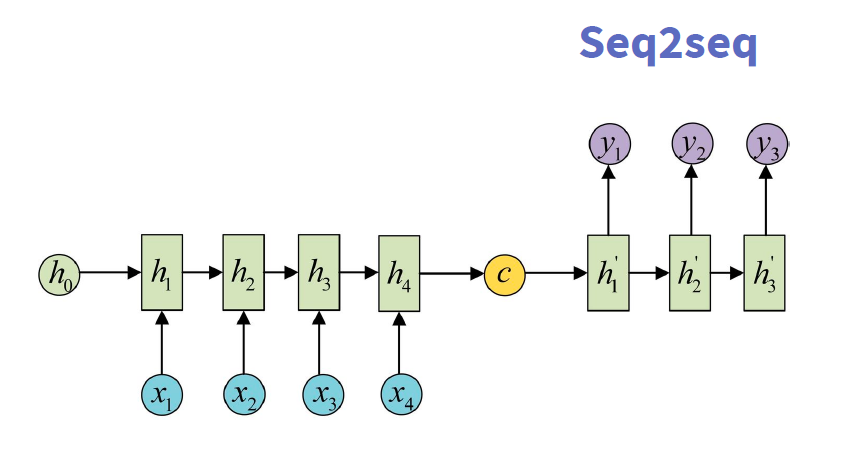


Text Rank的思想跟page Rank 类似，只不过它是将文章分为句子，然后转换成句子矩阵向量，句子矩阵向量转换成图，然后求取矩阵中的权重排行，从而矩阵中的权重对应的则是文章中的句子。

当我们在用Text Rank做抽取任务时主要步骤



生成式的方法主要有seq2seq



其中C代表前面输入信息经过一个RNN获得的一个context,然后将提取context再经过一个RNN进行对应的信息转换。

Seq2Seq可以用于图片视频等进行信息抽取，也可以用来进行翻译。著名的Encoder2Decoder就是代表。

它的整个思想是

将 输入 传送至 编码器，编码器返回 编码器输出 和 编码器隐藏层状态。

将编码器输出、编码器隐藏层状态和解码器输入（即 开始标记）传送至解码器。

解码器返回 预测 和 解码器隐藏层状态。

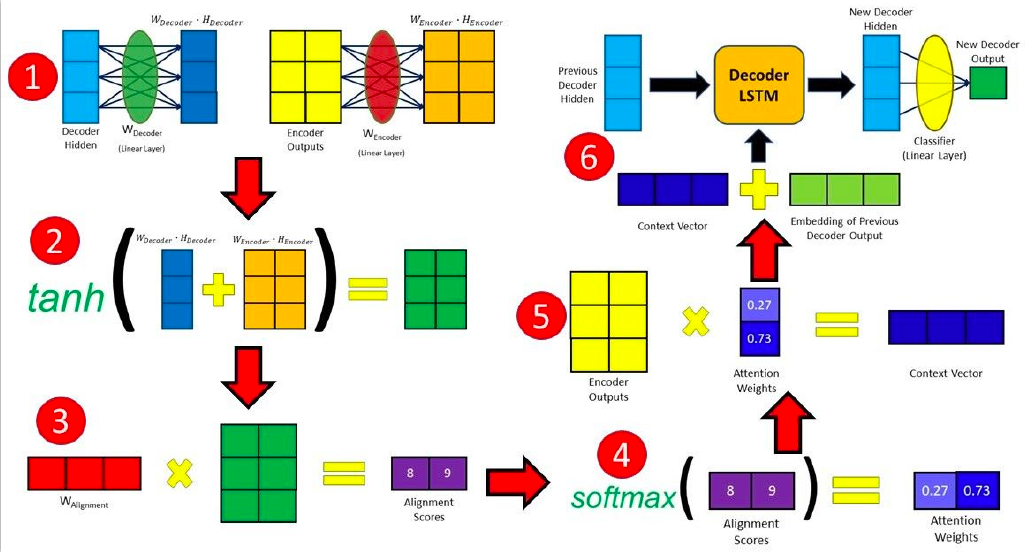
解码器隐藏层状态被传送回模型，预测被用于计算损失。

使用 教师强制 （teacher forcing） 决定解码器的下一个输入。

教师强制 是将 目标词 作为 下一个输入 传送至解码器的技术。

最后一步是计算梯度，并将其应用于优化器和反向传播。

但是我们进行Decoder时，在考虑我们解码的内容不希望与编码的所有都是相关的，可能是与其中一些是极度相关的，而与其他的是极度无关的，这就引入了attention机制，它的思想是将encoder的内容进行权重分配。Attention的机制有多种，其中包括了加法，乘法等。下面主要介绍加法



可以从上图看到它的计算方法，加法是将encoder的hidden和权重矩阵相乘，然后与decoder中的向量与权重矩阵相乘的结果相加。然后归一化就得到了encoder中的权重矩阵，然后将权重矩阵与encoder的输出相乘与之前的decoder的output向量相加就变成新的context，传递给decoder部分。乘法的区别在于它权重矩阵不是相加得到的，它是encoder的hidden与decoder的hidden相乘，然后归一化得到的。

当我们进行decoder时候也是需要输入的，我们一般可以先将一个Ground Truth作为输入，然后decoder得到的输出作为下一个输入。但是我们训练的时候是不知道Ground Truth，如果得到的输出作为下一个输入是错误的信息，则整个训练效果变得很差。解决方法是通过投硬币的方式决定是否采用上一个真正的输出还是随机一个作为输入。