《End-to-end Neural coreference resolution》论文阅读笔记

发布时间：2017 EMNLP

项目代码：https://github.com/kentonl/e2e-coref

主要工作：提出了第一个端到端的共指消解模型。核心思想就是先找出document中的所有潜在mention，然后对于他的所有可能的前指计算一个分布，优化目标则是gold coreference clusters中所有前指的似然概率。

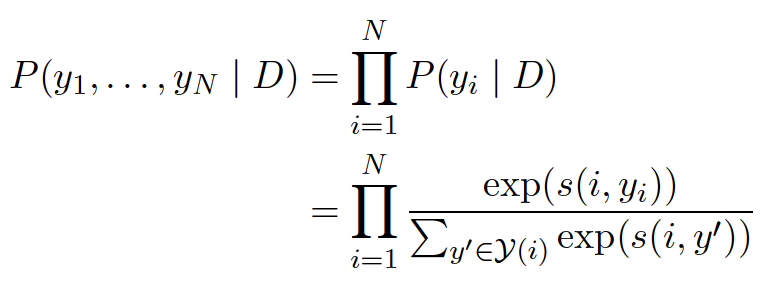
模型中包含一个span-ranking模块来决定对于每个span哪个前面的span是较好的antecedent，其中模型的核心模块则是span的embedding represent，它联合了上下文的表示和一个head-finding attention表示。

**模型**

假设document有T个词，那么就会有N=T(T+1)/2个可能的span存在，那么这个任务就可以描述为对于每一个，都需要给他确定一个前指，也就是代表一个虚假的前指，对应着两种情况：

1. 这个span不是一个mention
2. 这个span是一个mention，但是他没有任何的前指，也就是说它是这个实体的第一次出现

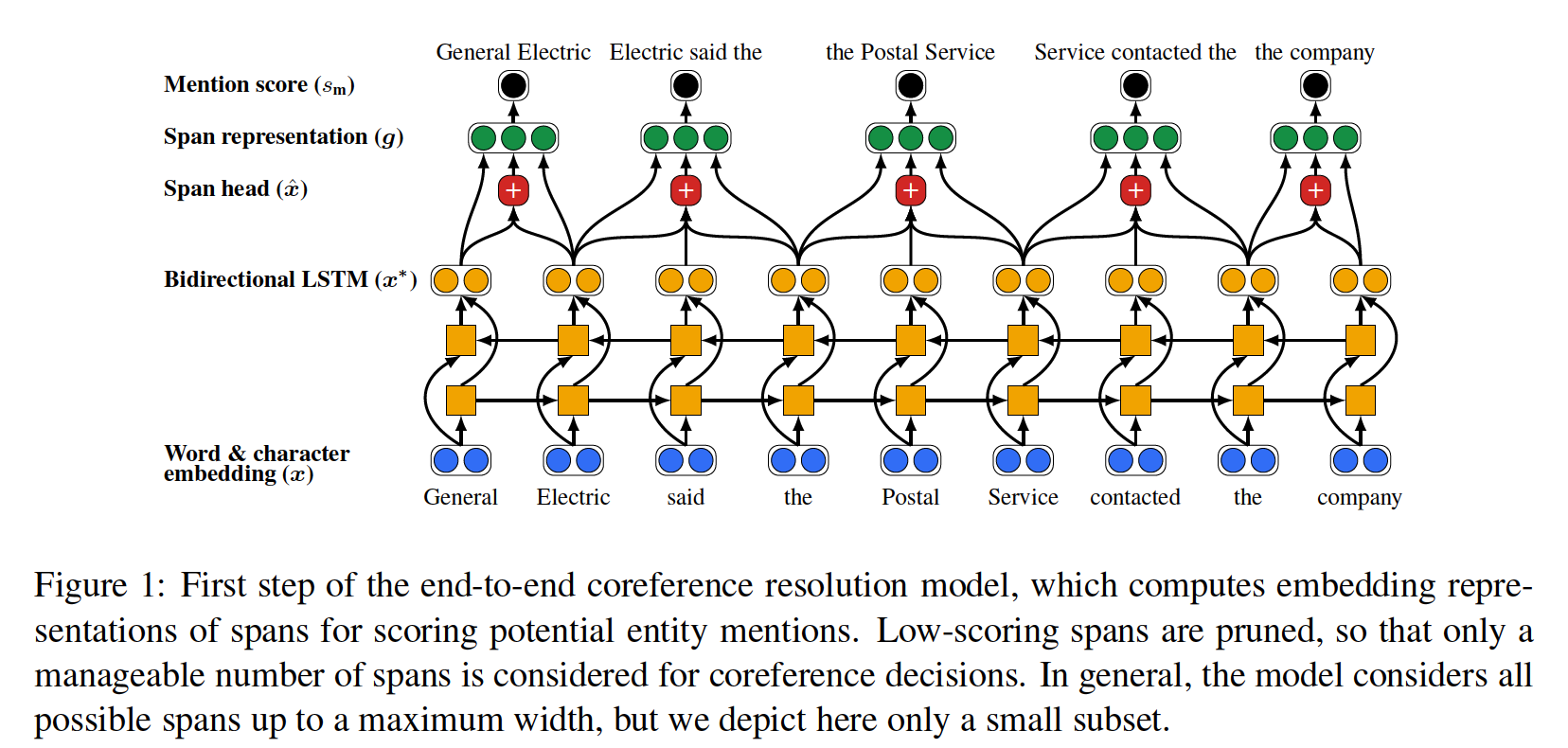
作者先计算每一个span以及他之前的span的共指得分s(i, yj)，然后将所有的span乘积得到一个条件概率分布，作者希望能得到正确的共指对。



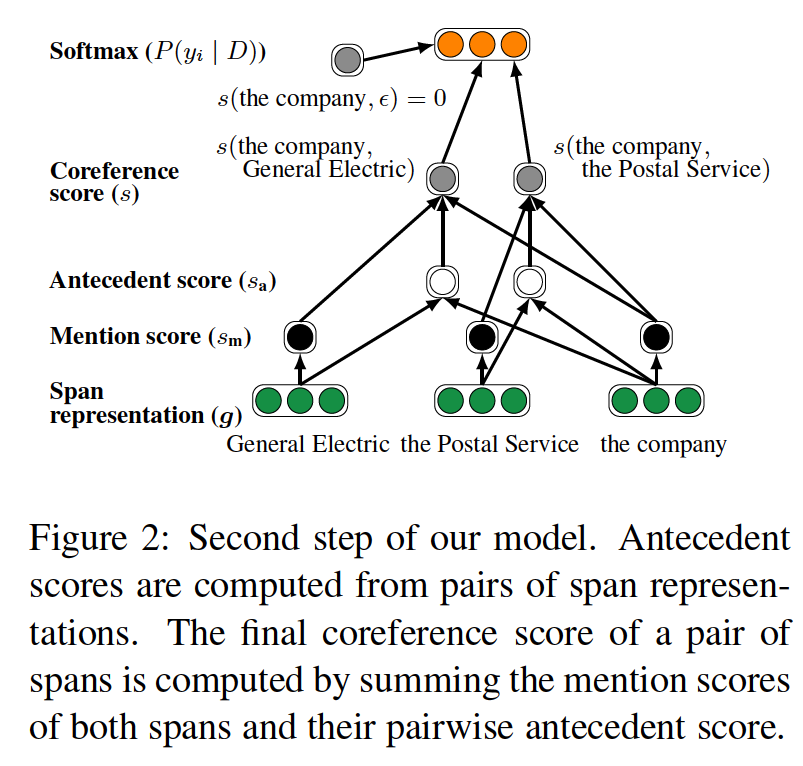
其中就是一个共指对的得分，它主要由三部分组成：

1. ：span i 是否是一个mention的得分
2. span j 是否是一个mention的得分
3. span j 是否是span i 的前指的得分

模型整个框架如下：



模型的第一步是对于所有span计算一个新的embedding表示，然后计算每个span成为mention的得分，把得分低的过滤掉。



模型的第二步则是对每个span对计算coreference score，而coreference score由两部分组成：mention score 和 antecedent score，然后加上虚假前指用softmax进行归一化。

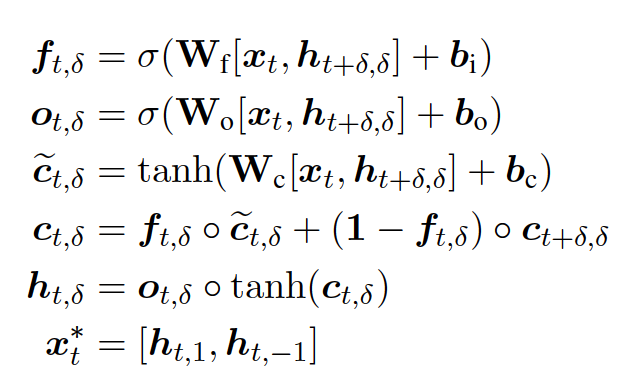
图1的embedding由两部分组成，一个是word embedding，一个是char embedding经过1维卷积后的结果。

**Span Representation**

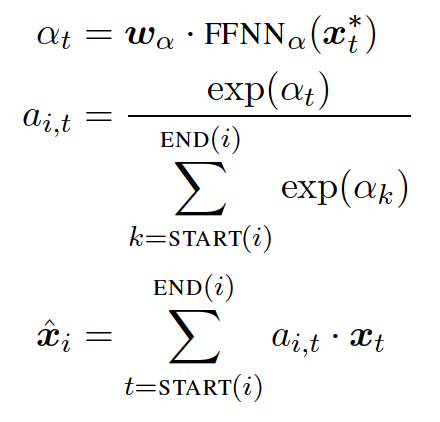
对于一个span的表示，最重要的有两点：

1. 这个span的上下文是什么
2. 这个span自己内部的结构

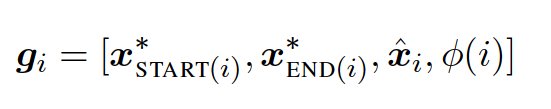
图1的embedding由两部分组成，一个是word embedding，一个是char embedding经过1维卷积后的结果。所以整个span representation在经过word embedding后，再经过一个双向LSTM得到新的表示，再利用attention机制得到新的span head表示，然后用这两个表示以及一些特征得到最终的span表示。



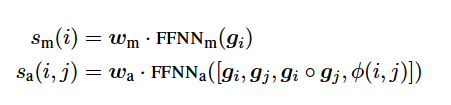
公式：LSTM的表示



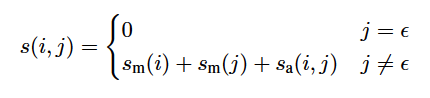
公式2：attention的表示



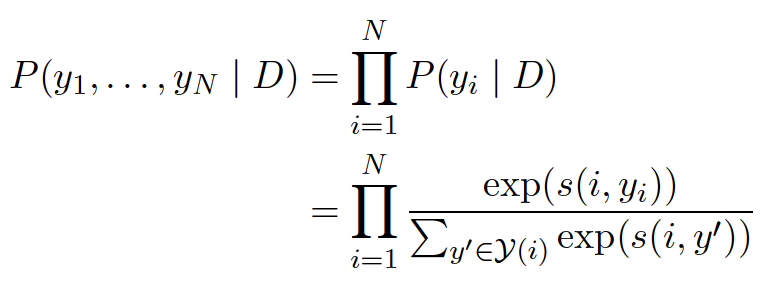
公式3：最终的span表示



公式4-计算前指得分和mention得分



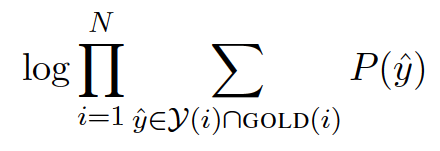
公式5--得分的计算方式



公式6--所有mention的概率

作者在实验时，还设置了一些额外的参数，一个span最大长度为L个word，只考虑分值最高的T个span，每个span最多只计算K个前指。

由于在训练时，我们只有共指簇的信息，所以作者希望优化gold cluster中的所有正确前指的边缘对数似然：



其中GOLD(i)表示包含span i 的共指簇，如果一个span i 不属于任何一个gold cluster，那么它的。

通过优化这个函数，模型自然会学到过滤掉那些不是前指的span。

作者介绍通过将虚假前指的分数置为0可以对模型中的mention检测做更好的约束，同时可以防止过滤span时引入噪声。比如，如果spani的唯一一个正确前指被过滤掉了，所以，在这个过程中，学习目标就会只降低非前指的分数，并且不会提高虚假前指的分数。

实验结果如下

作者报告了在数据集CoNLL-2012上的MUC等指标：

