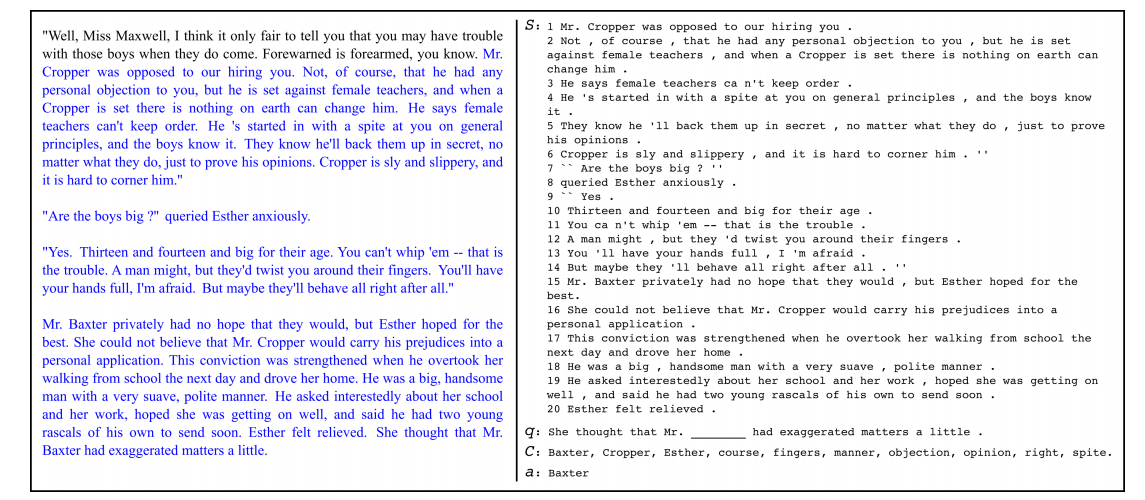
2019-07-18

基于完形填空的阅读理解（上篇）

* 1. 数据集：
     1. CBT（Children’s book test） dataset:

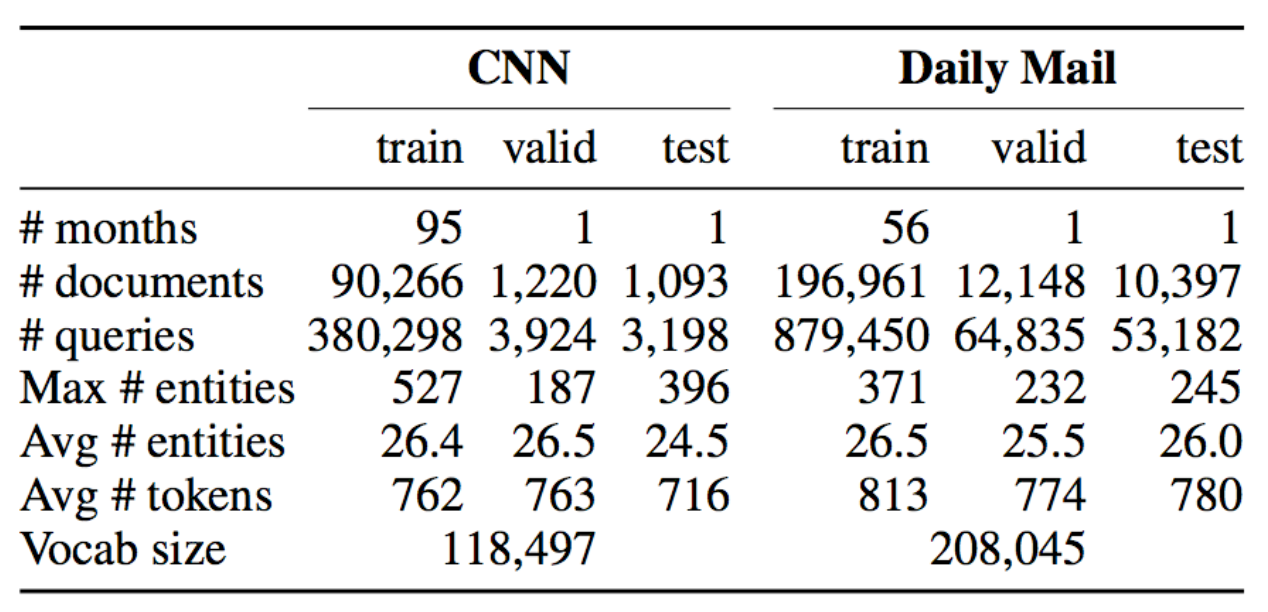
669,343 training questions  
8,000 dev questions  
10,000 test questions



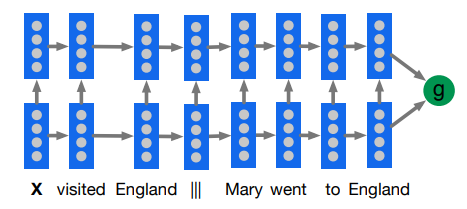
该数据集首先是一个文章，然后是从文章中抽出的前20个句子，接着是一个有空格的待填句子，接着是9个待选单词。

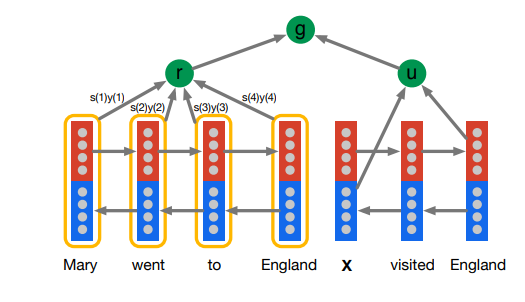
* + 1. CNN&Daily Mail数据集

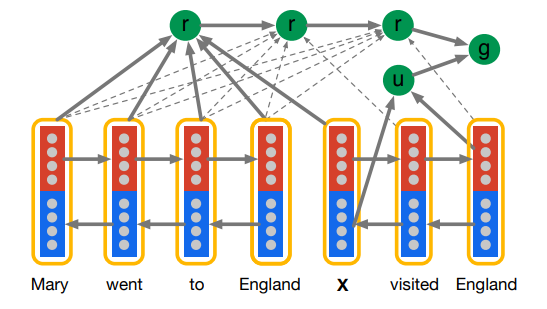
该数据集是从美国有线新闻和每日邮报中收集的大约一百万条新闻数据作为机器阅读理解语料库并通过实体检测等各种方法将总结和解释性的句子转化为[背景，问题，答案]三元组的数据集。其语料库中的信息内容如下：



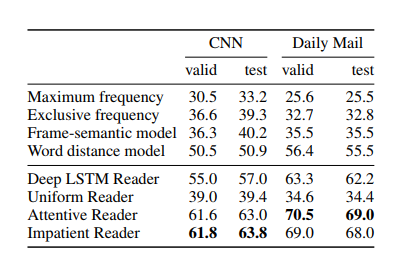
* 1. 相关工作：
     1. Hermann K M, Kocisky T, Grefenstette E, et al. **Teaching machines to read and comprehend**[C]//Advances in neural information processing systems. 2015: 1693-1701.







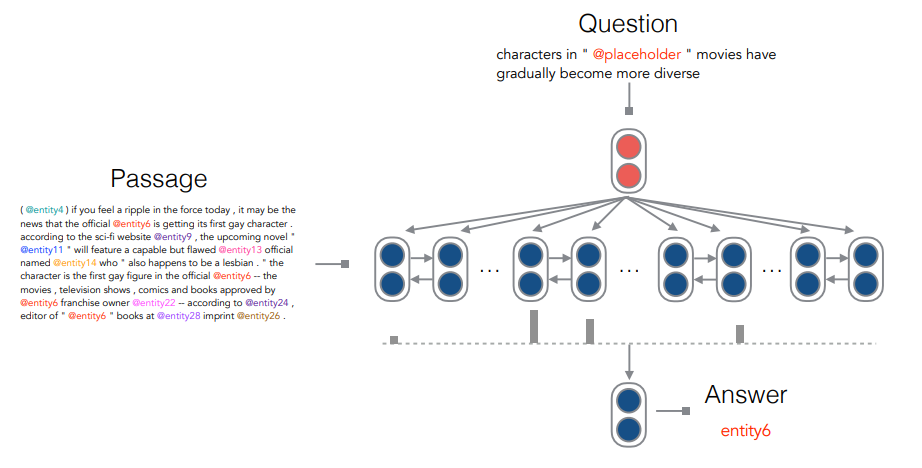
从上到下三个图分别为文章提出的deep LSTM reader, attentive reader和impatient reader模型，其结果如下：



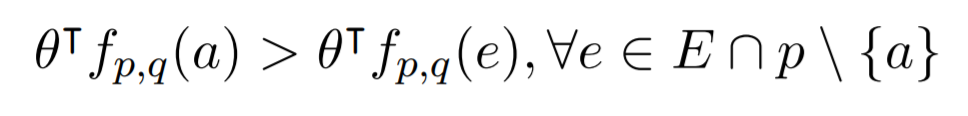
数据集使用了CNN和Daily Mail两个数据集，可以看出impatient reader 和attentive reader模型都取得了较好的结果。其使用的attention机制和重读机制很符合人类阅读习惯，可能因此取得了较好的效果。

但该模型在运算时输出了全词表的softmax，这显然是不科学的。

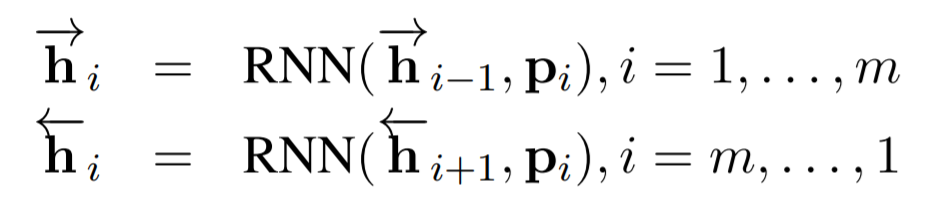
* + 1. Chen D, Bolton J, Manning C D. **A thorough examination of the cnn/daily mail reading comprehension task**[J]. arXiv preprint arXiv:1606.02858, 2016.



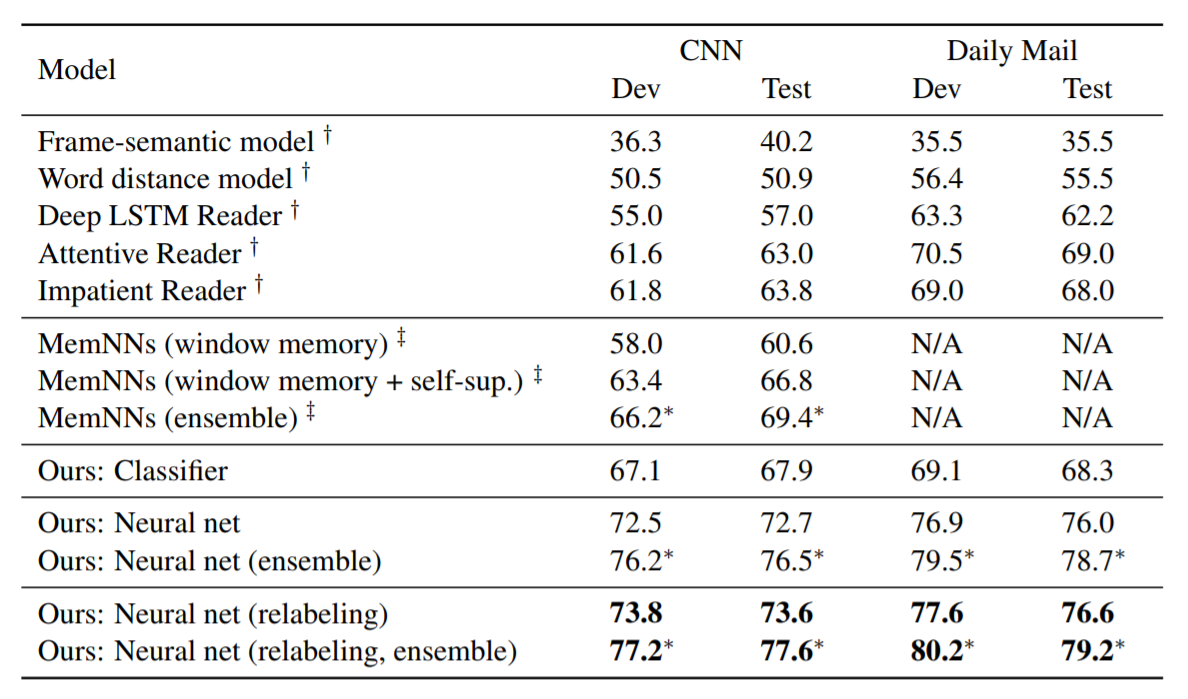
该文章分为两个部分，第一部分是构建了一个以实体为中心的卷积分类器，第二部分则构造了一个端到端的神经网络。卷积分类器每个候选实体构造一个特征向量，然后通过学习权重矩阵θ来为答案找到分数最高的候选实体，公式如下：



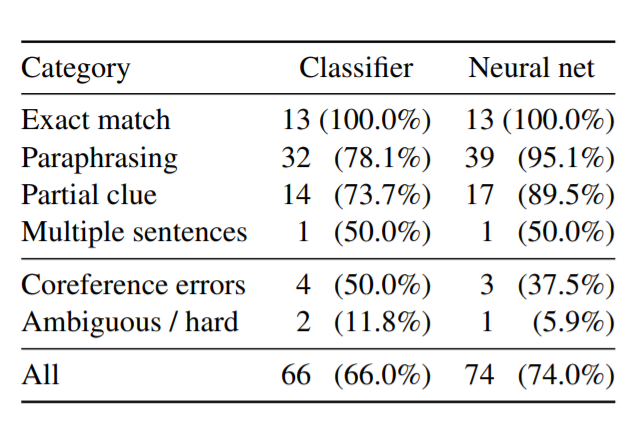
而端到端的神经网络结构如上图所示，文章将问题的编码和RNN的隐状态联合起来，如下图：



下一步，文章对比问题的embedding和所有上下文表示的embedding，并寻找与问题相关的信息片段。而后，根据计算得到的attention预测最可能填入其中的实体名称。文章结果分析如下：



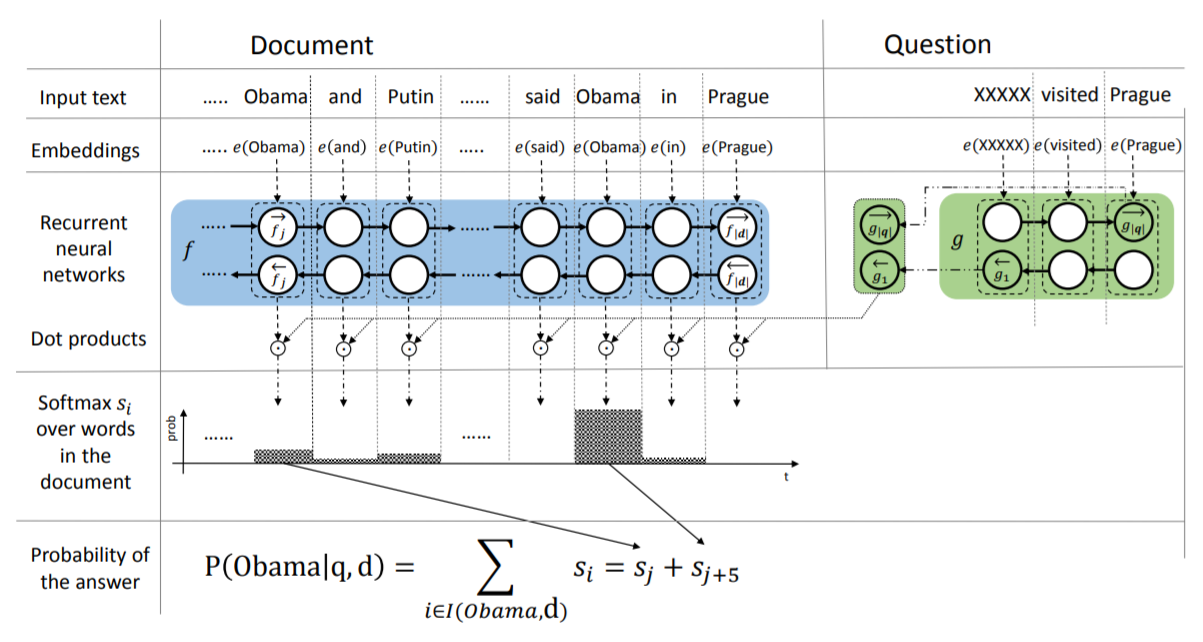
文章中对于多重语句，局部特征，及模糊不清的内容由于种种原因，表现不佳。



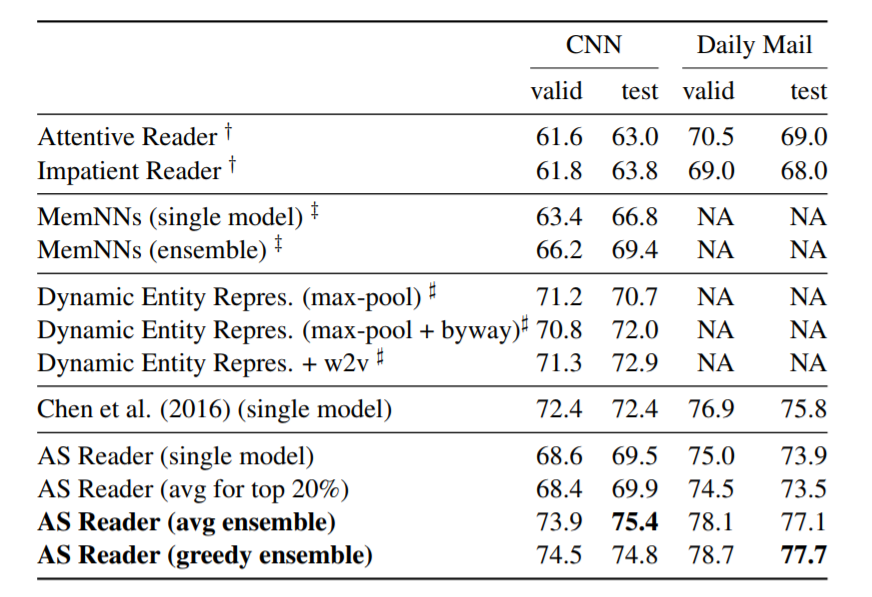
文章也指出，该数据集噪声较多，并且文章已经到达了相似结构的天花板。

* + 1. Kadlec R, Schmid M, Bajgar O, et al. **Text understanding with the attention sum reader network**[J]. arXiv preprint arXiv:1603.01547, 2016.

该文章是一个关于point network的文章，其结构如下图



文章在一开始就指出，虽然该模型得到了同时期最好的结果，但其无法预测一个没有出现在文章中的词，因为其输出的是整个篇章的所有词的softmax。算法共分为三步：1）计算问题的embedding向量；2）计算文章中每个词的embedding；3）使用问题的embedding和上下文embedding之间计算点积来计算文章中每个词是答案的可能性。同时，计算这种结果时，需要将文章中同一个词每次计算的概率加起来，作为最终的结果。其结果分析如下：



可以看出该算法对于两个数据集的提升非常之大。