英语作文自动打分综述

本文是对英语作文自动打分系统的(Automated Essay Scoring)综述。整个系统的流程如下：

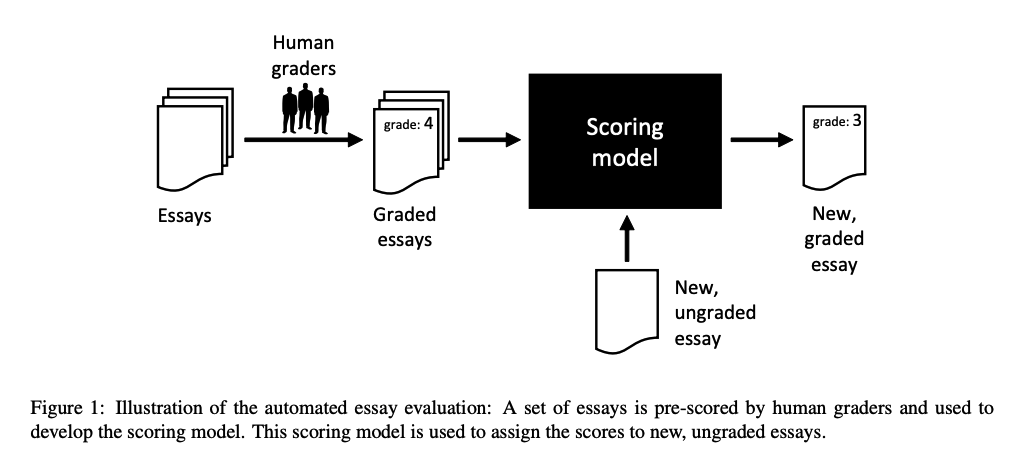


图1：AES系统流程

1. **PEG系统**

PEG（Project Essay Grade）系统是1966年由美国杜克大学的Ellis Page等人开发的。PEG的设计者们认为，计算机程序没有必要理解作文内容，在大规模考试中尤其如此。他们认为作文质量的诸多要素是作文的内在因素，无法直接测量，因此，最合乎逻辑的方法是从作文文本中提取一些能够间接反映作文质量的文本表层特征项（surface features）。

整个系统的流程如下：

* 收集一部分人工评分的作文（训练集），从作文中提取出若干个文本特征项（text features），并在这些文本特征项与人工评分之间进行相关性分析。
* 把文本特征项作为自变量，人工评分作为因变量，进行多元回归分析，得到回归方程。回归方程为每一个变量确定beta值。
* 在对新的作文进行评分时，PEG系统只需要提取这些变量，并把beta值代入回归方程之中，就可以得出这些作文的得分。

因此，PEG实现作文自动评分有三个步骤：变量提取、多元回归分析和把多元回归得到的beta值代入计算机程序计算出作文得分。

1. IEA系统

IEA（Intelligent Essay Assessor）是一种基于潜伏语义分析（Latent Semantic Analysis）的作文评分系统。与PEG显著不同的是，IEA的设计者们在其网站上申明：“IEA是唯一能够测量语义和作文内容的程序”。**潜伏语义分析主要是分析文本的内容和学生作文中所传达的知识，而不是作文的风格或语言。**

**­­­­潜伏语义分析（Latent Semantic Analysis，LSA）既是一种理论，也是一种技术。这种理论认为，在文本中隐藏着一个潜在的语义结构，这一潜在的语义结构正是所有词汇的语义之和。**

潜伏语义分析是一种矢量空间模型，但与一般的矢量空间模型相比，这种技术能够通过减少维数的方法，来有效的过滤干扰信息，提取数据中的潜在语义结构。**潜伏语义结构可以通过一个词汇项-文档矩阵来代表，矩阵中的每一行代表一个词汇项，每一列代表一个文档，而矩阵的每一个单元格中填入对应词汇项在对应文档中出现的频数。**

**使用矩阵代表潜伏语义结构的好处。是对矩阵可以进行一种称之为奇异值分解（Singular Value Decomposition，SVD）的矩阵分解，通过该运算，原来的矩阵可以被分解为三个不同的矩阵。减少维数后，对这三个矩阵进行进一步的计算，可以重建一个矩阵。重建后的矩阵因为使用了较少的维数，与原来的矩阵相比，可以更好地代表潜伏语义结构。该矩阵既保持了原来矩阵中的最重要的语义联想关系，同时又排除了大量的干扰信息。**

**将潜伏语义分析用于学生作文自动评分时，待评分的作文与预选定的范文（训练集）被视为矢量，对矢量进行比较之后，可以得到每一篇待评分作文与范文在内容上的相似度得分（similarity score）。这个得分就可以视为机器评分。**

**2.1 潜伏语义分析(LSA)**

潜伏语义分析是用来比较不同篇章的语义相似性。LSA的基本假设是不同文件的词语使用形式存在着某种潜伏的结构，而使用一些统计的方法可以估量出这种潜伏性结构。

那么怎么做LSA呢？

LSA以大规模的语料库为基础，使用了一种完全自动化的数学/统计技巧来抽取和推断词际关系，这种技巧就是线性代数中的奇异值分解（Singular Value Decomposition, SVD）。在SVD里，一个m\*n的矩阵A被分解成三个矩阵：一个是m\*n的成分矩阵U，把原来的列的实体表示成派生的正交因子值的向量；另一个矩阵w是n\*n的；第三个矩阵V是一个包含有标度值的直角矩阵，维度为n\*n。

**这三个矩阵相乘就可以恢复到原来矩阵，如果我们在恢复时减少维度，就可以取得压缩语义空间的效果。**

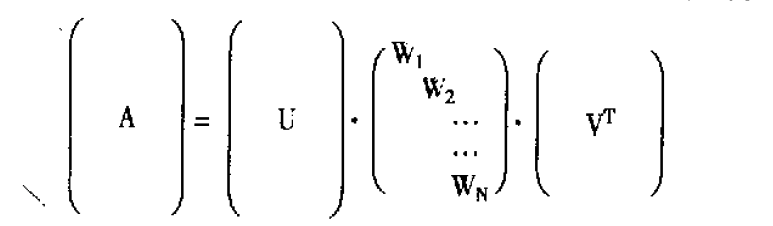
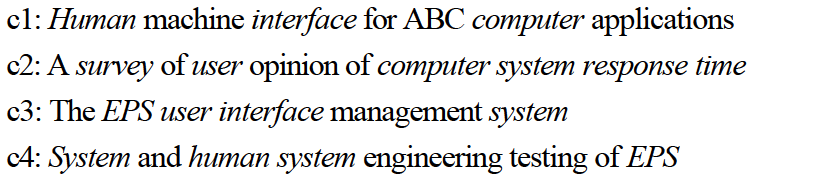
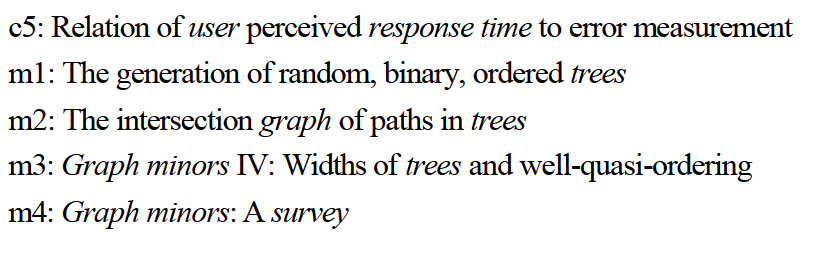
****

图2：SVD分解

比如下面的例子：总共有9条技术备忘录的题目，其中5条是关于人和计算机交互作用的，4条是关于数学图论的。这些题目在概念上不大相关，他们构成一个9列12行的矩阵，每一行是这些题目中出现起码两次的实义词，如下：

  
 图3：备忘录题目

统计结果矩阵如下：

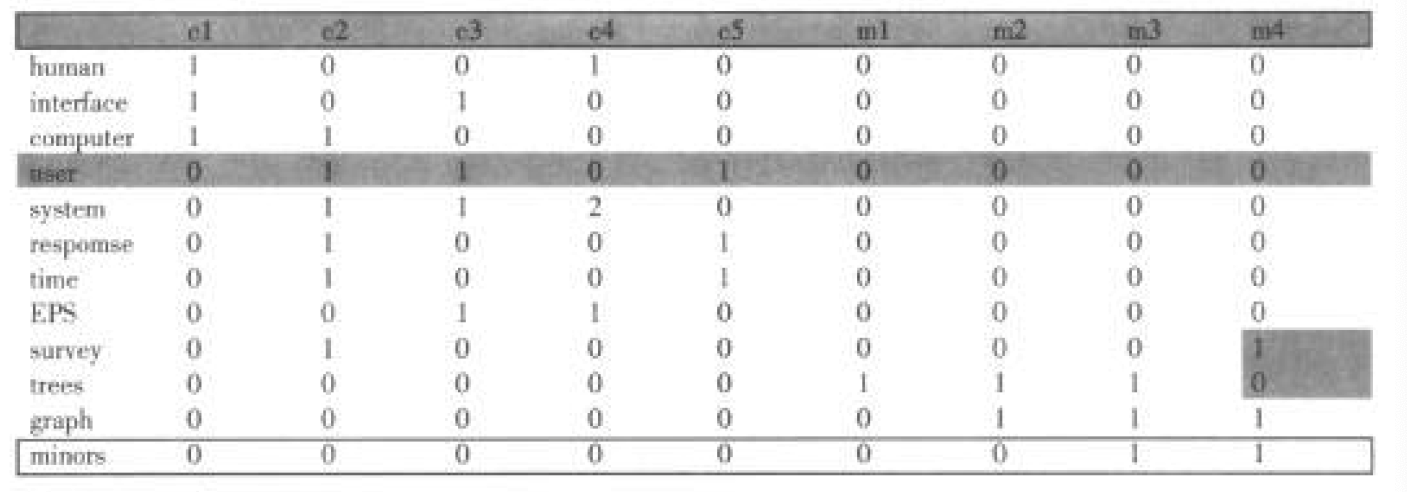


图4：统计矩阵

使用SVD技术，就得到下列的三个矩阵：

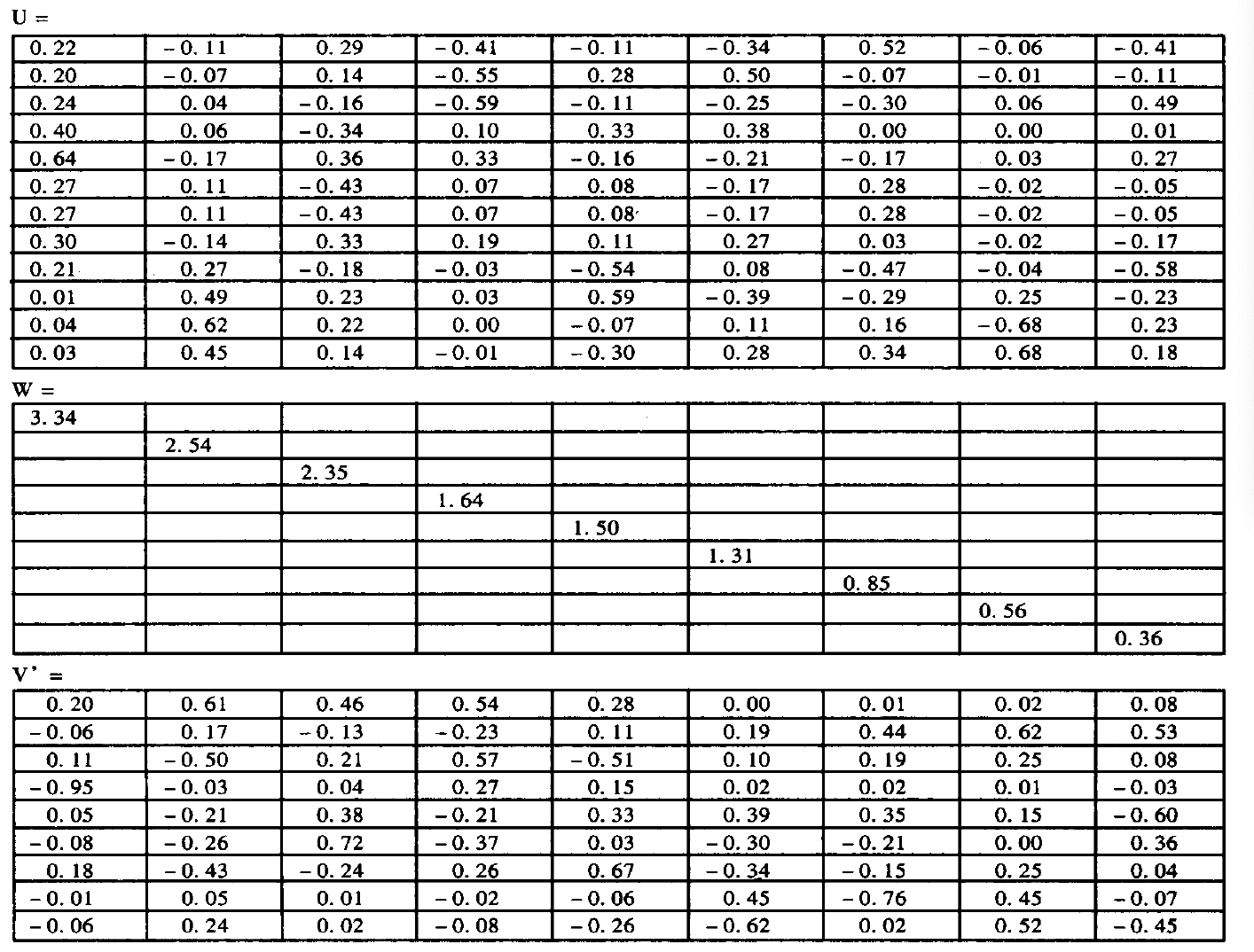


图5：SVD分解结果

其中U为12\*9的矩阵，W为9\*9的矩阵，V为9\*9的矩阵。

然后我们取这三个矩阵的头两列进行还原（U为12\*2，W为2\*2，V为2\*9）,这样我们就可以用三个较小的矩阵来表示原矩阵。

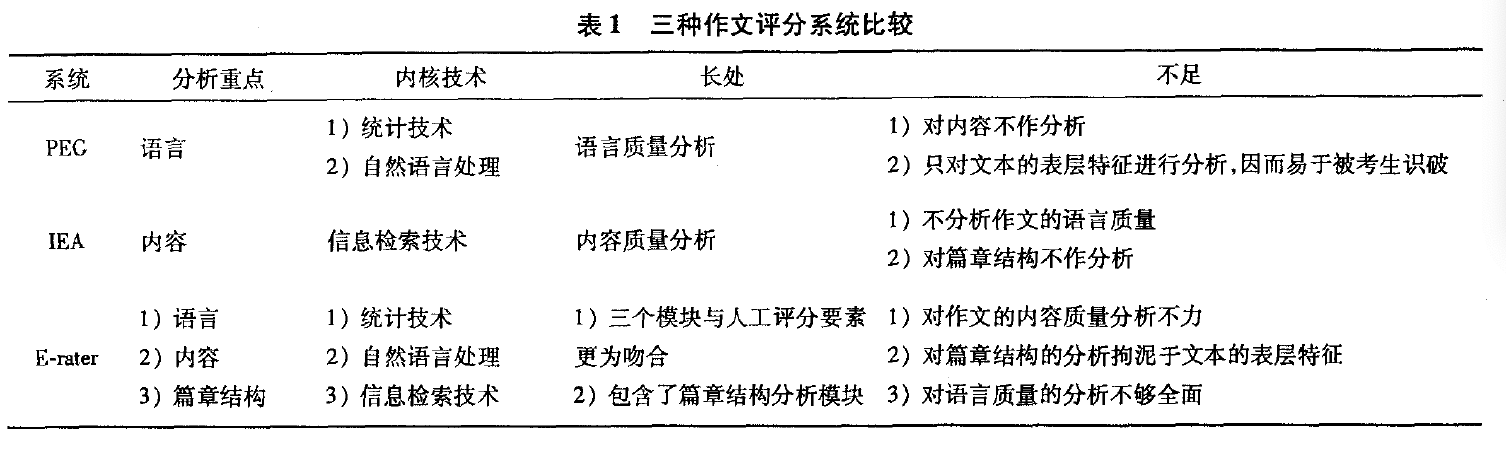
**三、E-rater系统**

E-rater系统是由美国教育考试处与20世纪90年代开发的，其目的是评估GMAT考试中的作文质量。E-rater的开发者们声称它不光能像PEG那样评判作文的语言质量，还能够项IEA那样评判作文的内容质量，除此之外，E-rater还对作文的篇章结构进行了分析。

E-rater的三个模块：

* 话语结构（篇章结构）：主要依靠在文中搜索“In summary”、“In conclusion”等提示词得以实现。
* 句法多样性：根据文中句子结构的多样性来评判作文的质量，显然该模块的目的分析作文的语言质量。
* 内容分析模块：在这一模块，E-rater通过矢量空间模型，观察作文是否包含了足够的与作文题目高度相关的主题词。

1. **三种系统的对比**

****

**图6：三种系统的对比**

1. **神经网络系统**

发布在2016ACL上的《A Neural Approach to Automated Essay Scoring》被认为是第一篇用神经网络来做AES。

他才用的数据集是Kaggle比赛的ASAP数据集，评分标准是QWK。

模型结构如下：

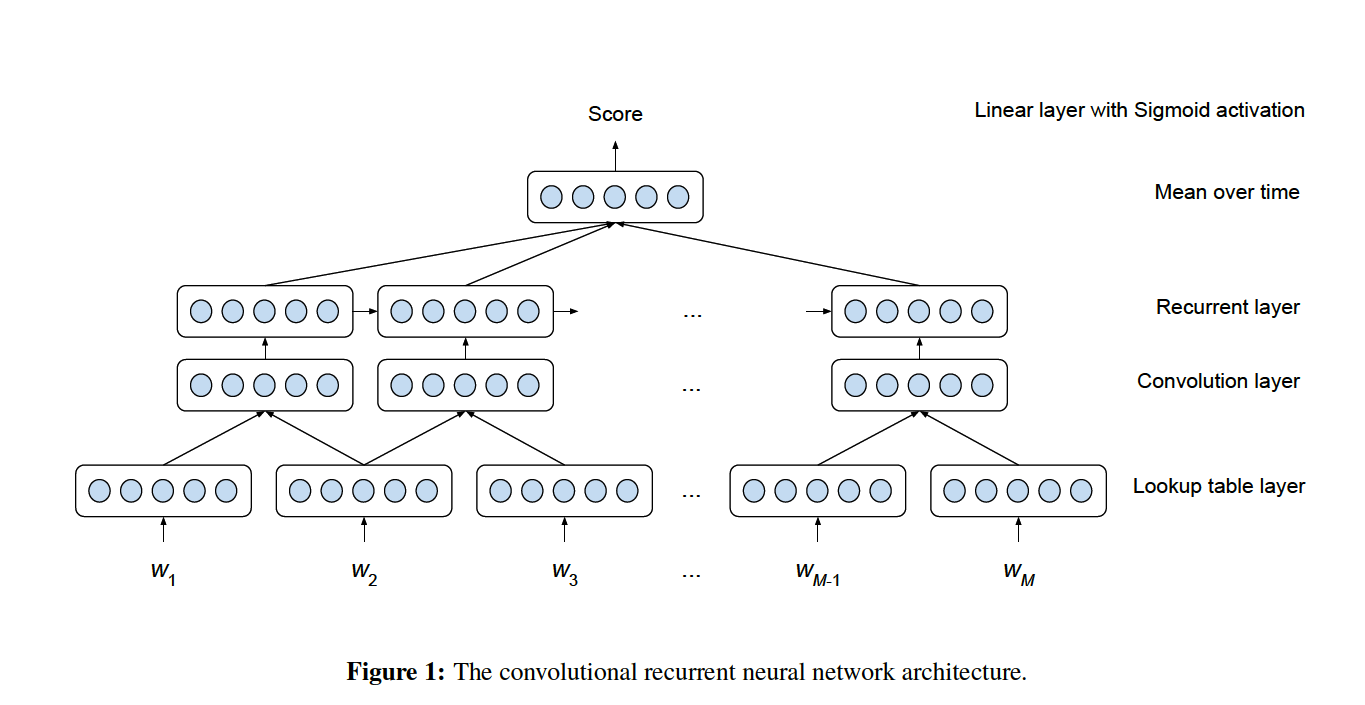


图7： 模型结构

**模型结构主要分为5层：**

* **Lookup table layer：也就是embedding 层，将词向量化。**
* **Convolution layer：采用卷积层，卷积过后维度不变，这里作者说卷积层可以提取文中的N-gram信息。**
* **Recurrent layer：循环网络层，采用循环网络，如果后续有Mean over time层的话，输出的就是每个隐层时刻的结果，不然就是输出最后时刻的结果。**
* **Mean over time：将上一层循环层的结果，每个隐层向量对位相加，得到一个向量表示。**
* **Sigmoid layer：采用sigmoid激活函数，将向量隐射为一个（0，1）之间的数值。**

**实验结果，作者采用Kaggle比赛的第三名公开的提取特征的代码，并用SVR（支持向量回归）和BLRR（贝叶斯回归）来训练模型。把他们作为Baseline。在构建神经网络模型时，采取不同的循环层(RNN、GRU、LSTM)，是否加CNN层与Mean over time层，以及模型融合，结果如下图：**

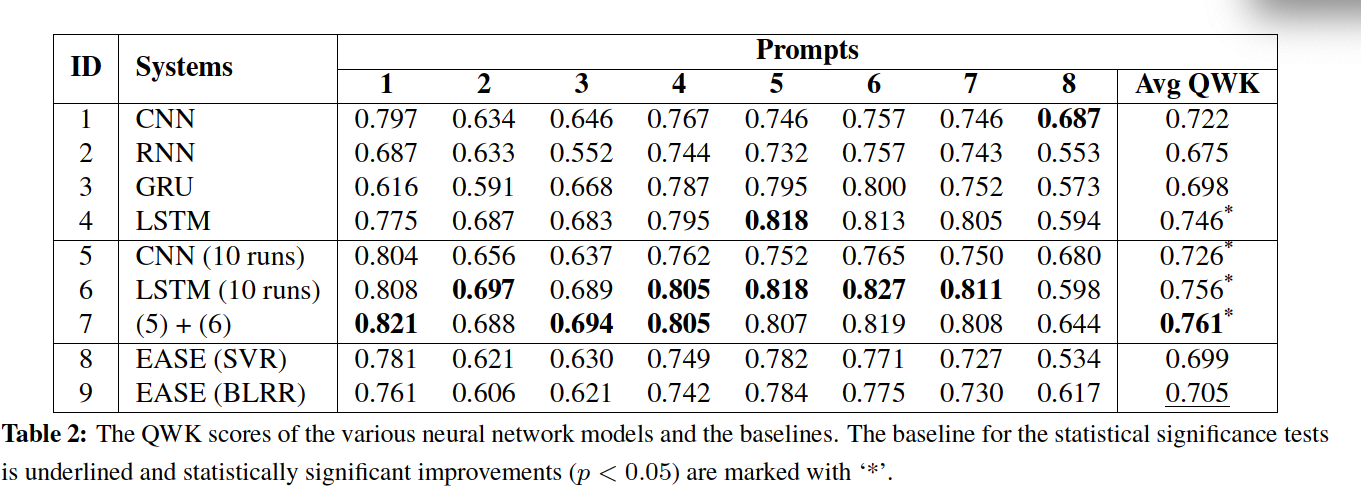
****

图8： 实验结果

从结果中可知，id 5的实验结果（ensemble模型）最高为76.1%，远高于baseline的70.5%。

后文待续。。。