**Word2vec训练中文词向量并生成《斗破苍穹》风格的自然语言**

**陈春祺**

学号:5140379022

Email:simoncqchen@outlook.com

**摘 要：**本论文基于谷歌提出的word2vec模型，提出了通过word2vec模型训练中文词向量的方法，实现了word2vec的skip-gram模型，以小说《斗破苍穹》作为语料库训练了中文词向量；此外，本论文还实现了基于RNN结构的语言模型，在训练过程中生成字向量，并生成与语料库具有相似语言风格的自然语言。

**关键词:** word2vec，词向量，RNN，自然语言生成

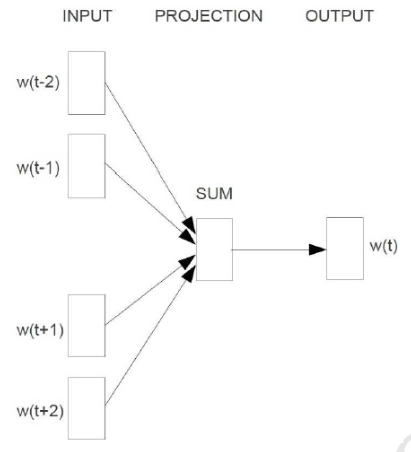
# 引言

# 相关技术

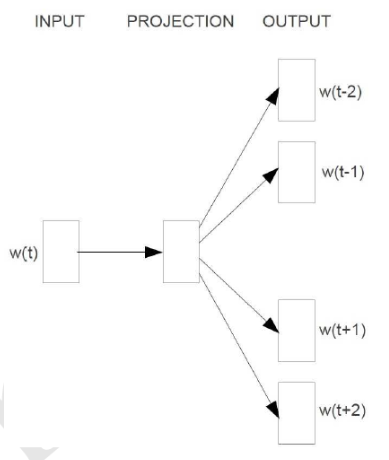
# Word2vec中文词向量模型

## Word2vec模型介绍

word2vec对词向量的训练有两种方式，一种是CBOW模型，即通过上下文来预测中心词（图3）；另一种skip-Gram模型，即通过中心词来预测上下文（图4）。其中CBOW对小型数据比较适合，而Skip-gram模型在大型的训练语料中表现更好。



**图3. CBOW模型**



**图4. Skip-gram模型**

CBOW模型包含三层：输入层、投影层和输出层。CBOW模型以上下文预测中心词，输入输出样本为，由w前后各c个词组成。模型的三层分别为：

输入层，包含中2c个词的词向量、、……、；

投影层，将输入层的2c个向量求和累加，如公式(1)所示；

输出层，对投影层的结果进行权重相乘、添加偏置值的计算，输出结果，作为对于中心词向量的预测结果。

类似的，skip-gram也包含三层，与CBOW不同的地方在于，由于skip-gram是从中心词预测上下文词语，输入层只有一个词向量，所以skip-gram的结构中的投影层是一个恒等投影，不对输入层的数据进行变换。之所以保留这样一个投影层是为了方便与CBOW模型进行对比。

## 中文词向量的训练

本文实现了word2vec中的Skip-gram模型。此模型对于中文词向量的训练步骤如下：

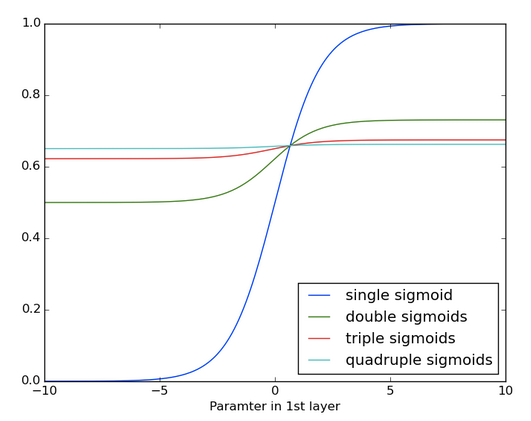
1. 使用Jieba分词，将原始语料库进行分词，按照词语出现的顺序将所有词语记录入一个列表。
2. 统计词频、词典等信息，将生僻词替换为UNKNOWN，将之前记录所有词语的列表转换为词语在词典中的序列的列表。
3. 构建skip-gram模型所需的训练数据。本文实现的模型为skip-gram，即根据中心词预测上下文，所以将记录所有词语序列的列表作为输入，而输入词语在原文中的上下文词语作为Label。具体来说，如果原文中存在“abcde”这句话，a、b、c、d、e分别为词语，上下文指中心词前后各一个词，那么根据这句话生成的输入数据与Label对如下：（b，a）、（b，c）、（c，b）、（c，d）、（d，c）、（d，e）。所有的训练数据生成完成后，再切分为minibatch。
4. 定义skip-gram的损失函数，在这里使用了tensorflow的nce\_loss函数，应用了负采样机制，提高训练速度并改善了最终训练结果中词向量的质量。使用梯度下降法来更新权值。
5. 对所有的输入数据进行训练，损失收敛之后记录当前的词向量作为结果。

# 基于RNN的语言模型生成自然语言

## RNN与LSTM

由于自然语言的特殊结构，在研究中很适合使用递归神经网络来建立语言模型。传统的语言模型包括n-gram语言模型和神经概率语言模型，和这两种语言模型相比，基于RNN的语言模型可以充分利用之前出现的所有词语的信息，并且处理的成本不会随着依赖上下文的增加而指数级增长。

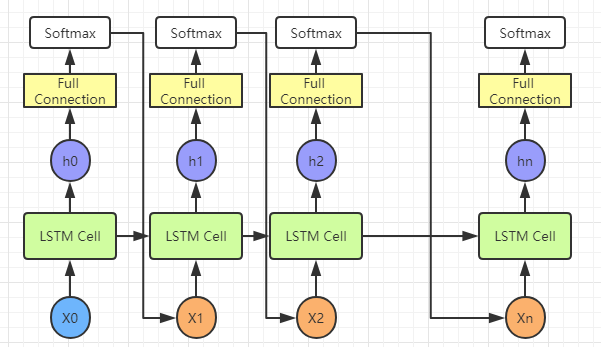
然而，在上世纪九十年代早期，梯度消失问题成为影响循环网络表现的重大障碍。正如直线表示x如何随着y的变化而改变，梯度表示所有权重随误差变化而发生的改变。如果梯度未知，则无法朝减少误差的方向调整权重，网络就会停止学习。原因之一是， 神经网络中流动的信息会经过许多级的乘法运算。由于深度神经网络的层和时间步通过乘法彼此联系，导数有可能消失或膨胀。梯度膨胀时，这些权重的梯度增大至饱和，亦即它们的重要性被设得过高。但梯度膨胀的问题相对比较容易解决，因为可以将其截断或挤压。而消失的梯度则有可能变得过小，以至于计算机无法处理，网络无法学习。这个问题更难解决。反复应用sigmoid函数的结果如图6所示。数据曲线越来越平缓，直至在较长的距离上无法检测到斜度。梯度在经过许多个层后消失的情况与此相似。



**图6. 多次应用Sigmoid后函数的曲线**

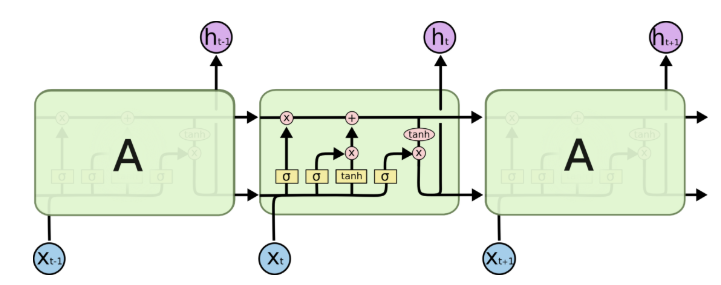
九十年代中期，德国学者Sepp Hochreiter和Juergen Schmidhuber提出了循环网络的一种变体，带有所谓长短期记忆单元，或称LSTM，可以解决梯度消失的问题。LSTM可保留误差，用于沿时间和层进行反向传递。LSTM将误差保持在更为恒定的水平，让循环网络能够进行许多个时间步的学习（超过1000个时间步），从而打开了建立远距离因果联系的通道。LSTM将信息存放在循环网络正常信息流之外的门控单元中。这些单元可以存储、写入或读取信息，就像计算机内存中的数据一样。单元通过门的开关判定存储哪些信息，以及何时允许读取、写入或清除信息。但与计算机中的数字式存储器不同的是，这些门是模拟的，包含输出范围全部在0~1之间的sigmoid函数的逐元素相乘操作。相比数字式存储，模拟值的优点是可微分，因此适合反向传播。这些门依据接收到的信号而开关，而且与神经网络的节点类似，它们会用自有的权重集对信息进行筛选，根据其强度和导入内容决定是否允许信息通过。这些权重就像调制输入和隐藏状态的权重一样，会通过循环网络的学习过程进行调整。也就是说，记忆单元会通过猜测、误差反向传播、用梯度下降调整权重的迭代过程学习何时允许数据进入、离开或被删除。在控制信息是否通过的过程中，LSTM很好的控制了梯度的变化，减轻了梯度消失的问题。基于这些现象，本课题的语言模型将基于RNN的变体LSTM来建立。

## 模型结构



**图5. 本文建立的语言模型**

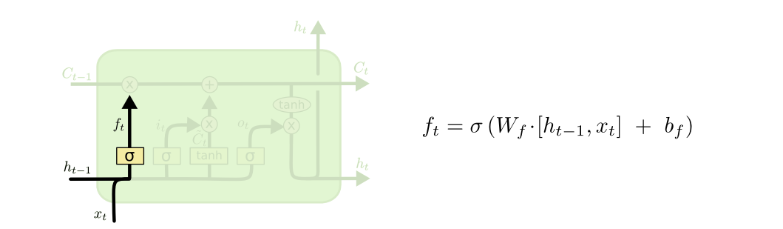
图5为本课题所建立的语言模型，模型初始的输入为x0，表示最初提供的字向量，文本的生成会从这个初始字符开始，经过LSTM Cell的计算得到当前的隐藏状态h0，将h0作为下一时刻的一部分输入；对h0进行一次全连接层的计算，经过Softmax层预测出下一时刻可能出现的字符，即x1，作为下一时刻的另一部分输入。循环这个过程，直到生成足够多的字符。



**图7. LSTM Cell结构**

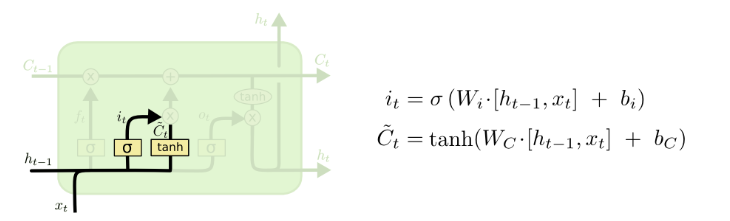
本文所应用的LSTM Cell结构如图7所示，当前LSTM的输入为、，LSTM Cell的计算过程如下：

1. 输入向量通过遗忘门，得到，计算公式如图8所示。



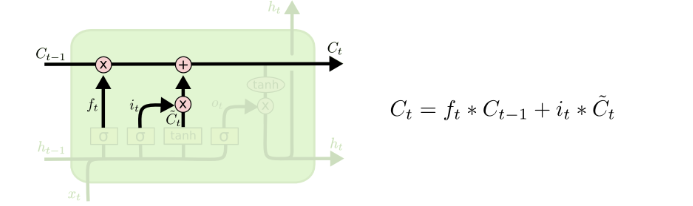
**图8. LSTM Cell遗忘门**

1. 这一步决定什么样的信息要被保留在当前细胞状态中，包括两个部分，一个输入门，进行sigmoid函数操作，决定将要被更新的值；一个tanh层，创建一个新的候选向量。计算公式如图9所示。



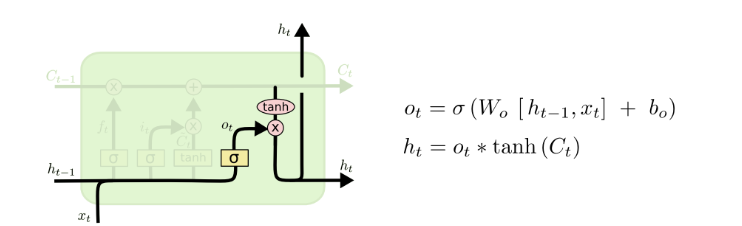
**图9. LSTM Cell输入门与tanh层**

1. 根据上一步生成的信息，更新上一时刻的细胞状态，公式如图10所示。



**图10. LSTM Cell更新细胞状态**

1. 决定需要输出的值，根据当前细胞状态和输入向量，通过sigmoid和tanh层计算得到隐藏状态，公式如图11所示。



**图10. LSTM Cell计算隐藏状态**

LSTM计算得到隐藏状态，对隐藏状态进行一次全连接层和Softmax层的计算（公式2），得到下一时刻输出字符的概率。

在训练过程中发现，如果直接选择概率最高的字符进行输出，生成过程会陷入循环，即不断生成之前出现过的句子，所以这里模型会选择概率大小前五的字符，将这五个字符的概率归一化，再根据它们归一化之后的概率随机选择其中一个字符作为下一时刻生成的字符，这个字符的子向量会和一起作为下一时刻的输入，重复进行计算，直到生成足够的字符，最后就是模型根据初始字符生成的自然语言文本。

# Word2vec中文词向量训练

## 数据集

本文所使用的自然语言语料库是小说《斗破苍穹》， 去除空格之后总词数3427356个，总字符数5260233；实验中使用的停用词表大小为1893。

## Word2vec词向量实验

本节描述了word2vec进行中文词向量训练的实验过程，实验中，训练参数的设置如表1所示。

**表1. Word2vec中文词向量训练参数设置**

|  |  |
| --- | --- |
| 参数名 | 值 |
| 词典大小 | 50000 |
| batch size | 128 |
| 词向量维度 | 128 |
| 上下文窗口大小 | 1 |
| 负采样个数 | 64 |
| 训练轮数 | 2000000 |

训练过程中损失随训练轮数的变化如图11所示。

**图11. 训练过程中损失随训练轮式的变化**

训练完后，选取了几个实例词语和词典中与它们最接近的词，如图12所示。可以看出，得到的结果是比较符合自然规律的，对于一些通用性的词语理解的很好，能比较准确的把握它们的语义，对于语料库中特有的词语，也能找到语料库中在语义上与其相似的词语。目前所使用的语料库还比较小，是一本单一的小说，如果增加语料库的大小，应该可以得到更好的结果。

**图12. 词向量和同义词的示例**

# RNN语言模型的训练和实验

信息、摘要和关键词

4.2.1 英文标题

英文标题置于论文第一页的最上方。主标题采用Times New Roman字体，居中，18磅，加粗，单倍行距，段前间隔0.5行。如有需要，可在主标题下方增加子标题，子标题采用Times New Roman字体，居中，11磅，加粗，单倍行距。

4.2.2 英文作者信息

英文作者信息置于英文标题下方。所有作者的姓名列于第一行，用逗号隔开。姓名采用Times New Roman字体，居中，10磅，加粗，行距12磅。

姓名下方放置作者的单位信息（英文），单位信息采用Times New Roman字体，居中，8磅，斜体，行距12磅。如果有多名作者并且单位不同，可以将不同的单位分多行编排，并用阿拉伯数字进行标注。

单位信息下方是作者的Email。Email信息采用Times New Roman字体，居中，8磅，斜体，行距12磅。如果论文提供多个作者的Email，可以用逗号隔开，并用阿拉伯数字进行标注。

4.2.3 英文摘要和关键词

英文摘要置于英文作者信息下方。摘要采用Times New Roman字体，10磅，行距12磅，两端对齐。

关键词至于英文摘要下方，采用Times New Roman字体，10磅，行距12磅，两端对齐，段前间隔0.5行。

4.2.4 中文标题

中文标题置于英文摘要和关键词的下方。主标题采用黑体，居中，18磅，加粗，单倍行距。如有需要，可在主标题下方增加子标题，子标题采用黑体，居中，11磅，加粗，单倍行距。

4.2.5 中文作者信息

中文作者信息置于中文标题下方。所有作者的姓名列于第一行，用逗号隔开。姓名采用仿宋体，居中，10磅，加粗，行距12磅。

姓名下方放置作者的单位信息（中文），单位信息采用宋体，居中，8磅，斜体，行距12磅。如果有多名作者并且单位不同，可以将不同的单位分多行编排，并用阿拉伯数字进行标注。

单位信息下方是作者的Email。Email信息采用Times New Roman字体，居中，8磅，斜体，行距12磅。如果论文提供多个作者的Email，可以用逗号隔开，并用阿拉伯数字进行标注。

4.2.6 中文摘要和关键词

中文摘要置于中文作者信息下方。摘要采用楷体（如有英文则使用Times New Roman字体），10磅，行距12磅，两端对齐。

关键词至于中文摘要下方，采用楷体（如有英文则使用Times New Roman字体），10磅，行距12磅，两端对齐，段前间隔0.5行。

4.3 正文

4.3.1 章节标题

章节标题可划分为三个级别。各个级别的标题均使用黑体（如有英文或数字则使用Times New Roman字体），加粗，行距16磅。标题用阿拉伯数字进行编号。

* 一级标题黑体加粗，12磅；段前、段后各空0.5行。
* 二级标题黑体加粗，11磅；段前、段后各空0.5行。
* 三级标题黑体加粗，10磅；段前空0.5行、段后不设置。

4.3.2 正文

正文使用宋体，10磅，各段落首行缩进2字符，两端对齐，行距15磅；如正文是英文格式的，行距为单倍行距，数字则统一为Times New Roman字体，并取消英文的“孤行控制”设置。

4.3.3 致 谢

致谢信息置于文章末尾和参考文献之间，致谢的采用一级标题的格式，12磅，黑体加粗，段前、段后各空0.5行，但是不使用阿拉伯数字编号。致谢的正文部分采用与文章正文相同的格式。

4.3.4 参考文献

参考文献的标题采用一级标题的格式，但是不使用阿拉伯数字编号。参考文献的标题使用英文（Times New Roman字体）和括号注明中文（黑体），加粗，字体为12磅，段前空1行，段后空0.5行，行距为单倍行距。

参考文献英文（Times New Roman字体），中文采用宋体，8磅，行距单倍行距磅，并采用“[x]”的方式以数字形式编号。

所有参考文献必须列出英文，也就是说，对于中文参考文献，必须先列出该文献的英文信息，在下方另起一行列出该文献的中文信息。

在正文中需要标注对参考文献的引用。标注时也使用“[x]”的形式，但采用上标格式[2]。

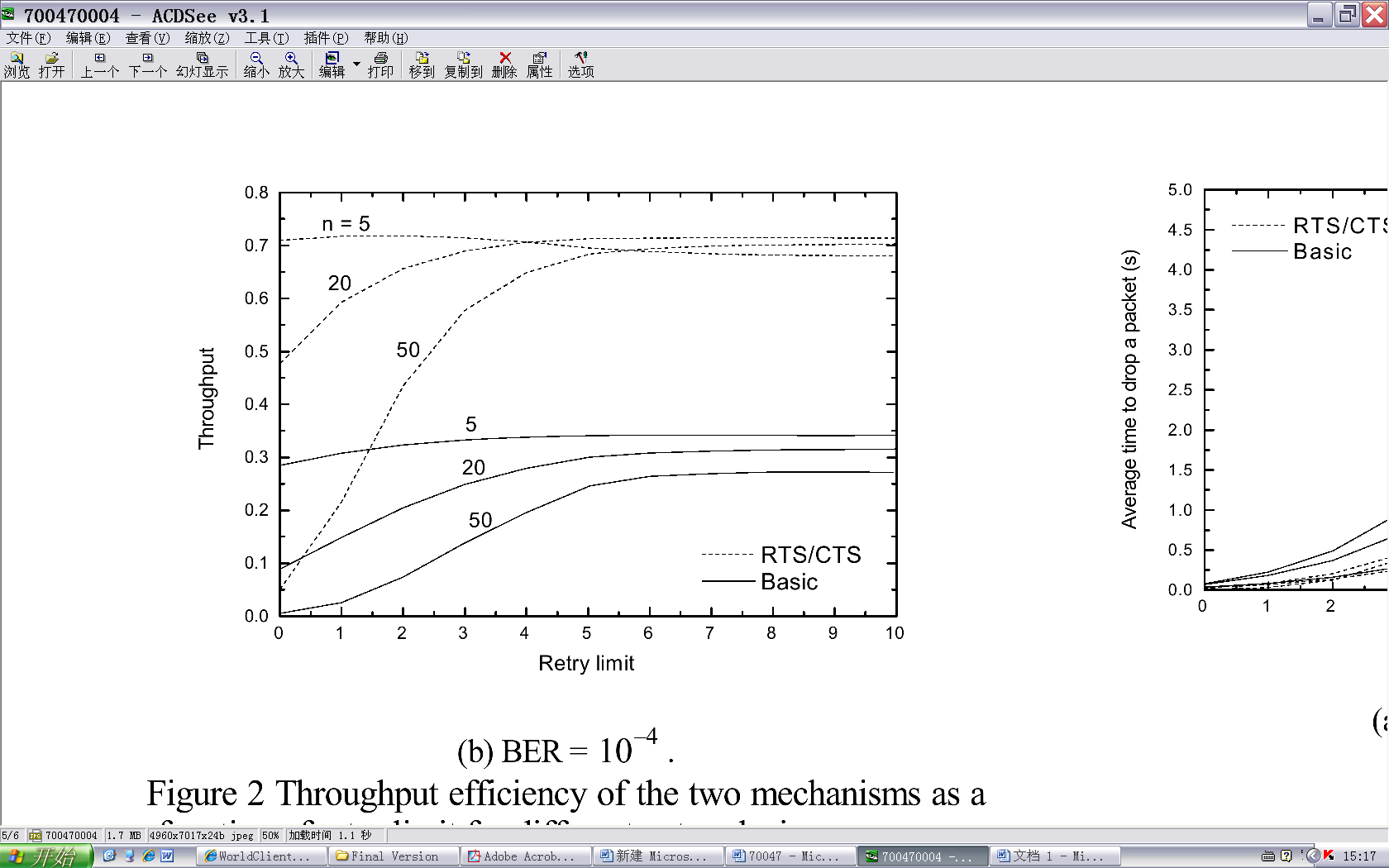
4.4 图、表和公式

4.4.1 图片

文中的图片应确保内容清晰。图片中的文字7.5磅。图片的尺寸可以根据需要适当放大或缩小，但是其长宽比例应与原图保持一致。对于比较大的图片，如果缩小后会导致内容不清晰，可以对该图片采用不分栏的格式。

所有图片应尽可能采用“嵌入式”环绕方式，尽量避免采用“四周型”环绕方式，否则排版过程中极易出现图片位置难以控制的情况。

图片居中。图片的标题放置于图片下方，所有图片必须列出英文标题，Times New Roman字体，如有中文采用黑体，8磅，加粗，居中，行距12磅，段前间隔0.5行，并使用“Figure x”（“图x”）的形式进行编号。图片的上方和图片标题的下方各设置一空行，行距15磅。



**Figure 1. Curve: system result of standard experiment**

**图1. 标准试验系统结果曲线**

4.4.2 表格

表格中的文字7.5磅。对于比较大的表格，如果按照双栏方式难以容纳，可以对该表格采用不分栏的格式。

所有表格应尽可能采用“无环绕”环绕方式，尽量避免采用“环绕式”。

表格应居中。表格的标题置于表格上方，采用英文Times New Roman字体（如有中文使用黑体），8磅，加粗，居中，行距12磅，段后间隔0.5行，并使用“Table x”“表x”的形式进行编号。表格标题的上方和表格的下方各设置一空行，行距15磅。

4.4.3 公式

对于嵌入在正文段落中的公式，如果因为正文段落15磅行距的设置导致公式不能完整显示，可以将该

**Table 1. System resulting data of standard expriment**

**表1. 标准试验系统结果数据**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 数量 | 质量 | 排序 | 稿件 | 件数 |
| 1 | 3 | 4 | 7 | 8 |
| 2 | 4 | 3 | 5 | 9 |
| 3 | 7 | 6 | 7 | 8 |
| 2 | 4 | 3 | 5 | 9 |
| 2 | 4 | 3 | 5 | 9 |
| 2 | 4 | 3 | 5 | 9 |
| 2 | 4 | 3 | 5 | 9 |

段落的行距设置为“单倍行距”，公式设定为：标准10磅，符号5磅，下标/上标6磅，次符号10磅，下标/上标5磅。

对于单独占据一个段落的公式，通常建议采用居中设置，并在段前、段后设置0.5行间隔。但该规则并不是强制性的，对于公式较多的论文，作者可以根据情况适当调整对其方式和段落间距，以求美观。

为求美观，应注意公式中的字体大小。字体过大会导致比例失调，字体过小会导致看不清楚。

致 谢

本章节为作者提供“致谢”的示例。

References (参考文献)

1. MENG Xiangping GAO Yan. Electric systems analysis [M]. Beijing: Higher Education Press, 2004. 3-21.
2. Li Yu, Liu Jingsen, Mechanism and Improvement of Direct Anonymous Attestation Scheme[J], *Journal of Henan University*, 2007, 37(2), P195-197 (Ch).

李煜，刘景森，直接匿名证言方案的实现机制与改进思路[J]，河南大学学报，2007, 37(2)，P195-197.