实验一：TTR 文本词汇丰富度分析

2021 年 5 月 22 日

学　　校:	华中农业大学
学院班级:	信息学院生信 1801 班
姓　　名:	邓启东
学　　号:	2018317220103
指导教师:	夏静波

目录

1 实验目的

1.1 TTR 反映文本复杂程度

类符与形符之比（type token ratio，简称 TTR）即一个文本中所有的类符（type）与所有的形符（token）的比值，其中形符指一个文本中所有的词数，类符指不重复计算的形符数，意思是在一个文本中，重复出现的形符只能算作一个类符。

通过计算文本的 TTR 值来判断一个文本的复杂程度，如果一篇文章用了很多不同的单词来写作，那么 TTR 的值就会接近于 1，说明这篇文章的用词比较复杂，词汇丰富度较高，阅读难度较大，反之，词汇丰富度越低，阅读难度越小。

2 实验材料

任一需要处理的文本或者语料库。本次实验的材料为 GENIA 语料库和 AGAC 语料库。

2.1 语料库介绍

2.1.1 GENIA 语料库

GENIA 语料库是一个生物医学文献集合。这个语料库是为了发展和评估分子生物学信息检索及文本挖掘系统而创建的。GENIA corpus version 3.0. 包含 2000 篇来自 MEDLINE 数据库的摘要。这些摘要是由 PubMed 按照 human、blood cells 以及 transcription factors 三个医学主题词（medical subject heading terms ）为搜索条件搜索到的。这个语料库已经被按照不同级别的语言信息、语义信息进行标注。

其包含：

1. 词性标注 2. 短语结构句法注释 4. 事件注释 5. 关系注释 6. 指代消解注释 (跨句结构里面，代词到底指的是什么) (GENIA 下载链接)

2.1.2 AGAC 语料库

AGAC（Annotation of Genes with Alteration-Centric function changes）是一个活性基因人类专家注释语料库，目的是捕捉突变基因在致病环境中的功能变化。可以用于药物再利用的案例研究，揭示了变异与广泛的人类疾病之间的潜在关联。AGAC 注释了 11 种命名实体类型。通过识别出 LOF/GOF 分类的基因-疾病关联，结合 LOF-激动剂/ GOF-拮抗剂假说可以应用于疾病候选和物质的知识片段的发现。(AGAC 下载链接)

3 实验步骤

3.1 全语料库 TTR（类符/形符比的计算）计算

我们首先想要分别统计两个文章所有语料的整体类符与形符之比。

GENIA 文本语料库文本由 GENIAcorpus3.02.xml 直接复制得到，并通过代码删去空行。GENIA 语料库由 2000 篇文献，每一篇文献占据三行，分别为 MEDLINE 文章 ID、文章标题、以及摘要正文。我们将第三行摘要正文提取并以列表存储（方便后续实验抽样），整合为一个大的文本文件 (见附件 TTR.py)，并上传至服务器使用 linux 命令行进行整个语料库的 TTR 计算 (见附件 linux 命令行.txt)。

AGAC 采取同样的方法 (代码见附件 AGAC_process.py)，处理数据过程中发现，AGAC 文章大多以两行式存储：第一行标题, 第二行摘要。但有 43 篇例外，并列举出其文件名称、是第几篇文献、以及其实际行数 (见附件 AGAC 非两行式文章.xlsx)。

其最终 TTR 计算结果如下表1所示。

TTR 结果			
预料库	分子	分母	TTR
GENIA	13,648	434,141	3.14%
AGAC	7,132	55,414	12.87%

表 1: 两语料库 TTR 计算结果

我们发现，原文件 GENIA_abstract.txt 的总单词数目为 402,694，而 GENIA_abstract.pure.txt 的总单词数目为 434,141。转化成 pure 后单词总数多了三万。分析原因是由于在转换成单个单词的时候，将连字符（hyphen）变成回车符导致复合词语被拆成了多个词语。

由于分子是基于拆分后的 GENIA_abstract.pure.txt 文件统计得出，因此分母应取 434,141。分子为 13,648。结果为 3.1438% ，同理 AGAC 的 TTR 值为 12.87%。

由计算出的 TTR 可以发现，AGAC 的结果 TTR 较大。是否就意味着 AGAC 语料库的文本词汇丰富度大，阅读难度大呢？

实际上不是。只有在篇幅相同或者相近的时候，TTR 越高才能意味着用词越丰富。但当篇幅差异较大的时候，比如 GENIA 的语料库大小远大于 AGAC，近十倍的情况下, 不可直接对比 TTR, 这是因为词汇储备量及丰富度主要体现在实义词上, 但随着篇幅增加, 写作中不可避免的功能词 (如 the, a, of, and 等) 会不断重复, 这会稀释 GENIA 语料库整体的 TTR。

因此, 篇幅不一时要对 TTR 进行标准化, 计算每百词 (根据长度可调整为每千词、每万词) 的平均 TTR, 即 STTR（标准化类符/形符比）的计算方式。

3.2 语料库每 1000 词 STTR（标准化类符/形符比的计算）计算

3.2.1 STTR 的计算方式

标准化类符/形符比的计算方法是, 计算每个文本每 1000 词的类符/形符比, 将得到的若干个类符/形符比进行均值处理.（如某文本长 5000 字, 其中第一个 1000 词的类符/形符比为 50%, 第二个 1000 词的类符/形符比为 52%, 第三个 1000 词的类符/形符比为 54%, 那就把这三个数字平均下得到 52%）。因此我们需要对两个语料库已经分词的 pure 文件的单词进行随机抽样。

3.2.2 两个预料库 STTR 结果

经过反复的比较，我们发现在抽样次数达到 10000 次，计算每 1000 个单词的 STTR 的时候 (代码见附件 STTR.py)，两个语料库结果均稳定在近 52%(表2) 这和之前的 TTR 结果大为不同。首先，由于是计算每 1000 个单词的 TTR，数目较小，很多单词的重复率不高，整体的 STTR 都较大。并且，在选取的词数相同，即标准化以后，可以看出两个语料库的标准化类符/形符比大致相同，可以说明两语料库的词汇丰富度是相近的。

语料库	STTR
GENIA	51.9494%
AGAC	51.9634%

表 2: 抽样得到的语料库的每千词 SSTR

4 实验总结

4.0.1 TTR 的局限性以及标准化后的 STTR

我们从本次实验可以看出，使用 TTR 和 STTR 得出了两个完全不同的结论：TTR 计算出的 GENIA 和 AGAC 两个预料库的词汇丰富度相差 4 倍。显示两个语料库的差异极大。但是经过分析后我们考虑了语料库的大小因素，引入了改良版本的 STTR 这个评价指标。最终多次随机抽样取得的平均 STTR 结果十分相似。得出了完全相反的结论。可以看到评价指标在这个当中的重要性，因此一个实验我们要考虑到实验的合理性，其他因素有没有排除，才能保证我们的结论是正确的。

4.0.2 其他方法

L2SCA 是宾夕法尼亚州立大学的 Lu 教授于 2010 年基于 Python 语言下开发的句法复杂度分析器, 涵盖了 14 种句法复杂度指标, 可供句法研究者及教师使用. 但是基于其需要在 Python 语言编辑器中使用, 因此本文的目的是说明其使用方法, 以供更多研究者研究使用。

4.0.3 实验后的一些联想

实际上标准化的思想在 RNA-Seq 基因差异表达分析、词向量嵌入中都有所体现。在学习当中举一反三，我们可以提升自己独立思考的能力。