Education and Learning, 教育与学习(3)2021, 6
ISSN:2705-0408(P); 2705-0416(0)

基于Sentence-BERT的智能合同对比分析系统设计

罗　通
（三亚学院信息与智能工程学院　海南三亚　572099）

【摘　要】本文借助Sentence-BERT深度学习模型，利用相关算法，旨在提高文本类工作办公效率，利用技术手段对合同、论文和标书等文本进行对比分析得出对比结果，方便客户可以更加清晰的查看两份文档的异同，进行设计智能合同对比分析系统，以此解决诸如文本智能分析、标书串标问题、合同各个版本快速找出异同等一些社会实际问题。本系统设计的目的是为了减轻大量繁琐人工审查工作带来的负担，缩短查重的时间，减少学术造假带来的不良社会风气，有效帮助企业减少在项目投标过程中多家公司串标造成的经济损失事件的发生。

【关键词】Sentence-BERT网络结构；文本对比；文本相似度

DOI：10.18686/jyyxx.v3i6.47831

随着计算机技术的不断进步，文本查重技术也被广泛使用在多个应用场景，最为凸显的是学术论文查重。一般的查重系统中有一个对比库，上传进行检测的论文内容都会与对比库中的资料进行对比，来检测论文内容是否抄袭。但当前方法还存在一定的技术缺陷，如不能够通过语义来判断语句是否为相同意思。

现在我国的合同修改审阅等工作都需要大量的人力物力，虽然有一些机构拥有基本语句校验辅助工具，却没有成型的文本对比、智能分析这样的软件工具可以使用。不管是传统的纸质方法还是人工审阅方式，以及在这方面管理的模式，在该领域这种方式已较为落后，在人工智能和大数据技术飞速发展的今天，各行各业中许多人工体力劳动已经在潜移默化被人工智能所取代，可以预知，传统的文本管理手段和审批手段在不久的将来也一定会被计算机程序所取代。

1　关键技术

1.1　开发工具和MySQL

IntelliJ IDEA 作为最佳的 Java 开发工具，理所当然是本系统采用的开发环境之一。同时采用性能卓越的 Maven 来进行项目管理，其具有以下优点：

（1）简化项目依赖管理；

（2）提供统一的构建系统；

（3）方便与持续集成工具整合；

（4）支持多模块项目的开发。

在数据的构建、关联和管理等方面，本系统采用性能稳定的开源数据库管理系统 MySQL。其具有以下优点：

（1）性能卓越，服务稳定，很少出现异常宕机。

（2）开放源代码且无版权制约，自主性强、使用成本低。

（3）软件体积小，便于安装，维护成本低。

（4）支持多种操作系统和多种语言开发，具有丰富的 API 接口。

1.2　Sentence-BERT网络结构

在了解 Sentence-BERT 网络结构之前，先来了解 BERT 模型。BERT 模型具有以下两个特点：其一，这个模型有 12 层的深度，但并不是很宽（wide），中间层只有 1024，而 Transformer 模型中间层则有 2048。这种深而窄的模型在图像处理方面往往会比较好用一些；其二，MLM（Masked Language Model）旨在表达融合上下文的能力，这使我们可以预先训练深度双向转换器。掩蔽语言模型随机掩蔽一些输入符号，以便预测掩蔽语言的原始词汇，单词仅基于上下文，使用“掩码语言模型”的预训练目标来减轻上述单向约束。

虽然 BERT 模型已经在 NLP 各大任务中都展现出了强者的姿态，其具有很高的准确度，但在通过其在句子对回归任务（如语义文本相似性（STS））上最新表现发现：在做句子相似度时它需要将两个句子都输入到网络中，这导致了巨大的计算开销：在 10000 个句子的集合中找到最相似的一对需要用 BERT 进行大约 5000 万个推理计算（约 65h）。基于 BERT 的构造，在语义相似性搜索和聚类这样的无监督任务方面，它似乎不是很合适。为了解决问题，Sentence-BERT 对 BERT 模型进行了改进，使用孪生和三元网络，能够输出保有语义信息的句子嵌入。通俗来讲，就是借鉴孪生网络模型的框架，将不同的句子输入到两个 BERT 模型中（但这两个 BERT 模型是参数共享的，也可以理解为是同一个 BERT 模型），获取到每个句子的句子表征向量；而最终获得的句子表征向量，可以用于语义相似度计算，也可以用于无监督的聚类任务。改进后的 BERT 模型，对于同样的 10000 个句子的任务量，大约 5s

左右就可以完成计算，从 65h 缩减至 5h 左右，这是极大的改进。

2 设计目标

本系统主要用于个体用户在文本处理过程中进行的一些辅助类工作。预计实现目标有以下方面：合同智能分析对比；文本智能分析；文本求相似度；PDF 类文本内图像相似度；文本简单语法矫正。

3 相关算法

3.1 余弦相似度

余弦相似度算法主要是通过测量两个向量之间夹角的余弦值来计算其的相似度。余弦相似度算法通常用于正空间，所以计算结果的区间在 0 到 1 之间。其数学表达式为：

$$\cos\alpha = \frac{A*B}{\|A\|*\|B\|} \quad (3\text{-}1)$$

拿到文本首先对两段文本 A 和 B 进行分词，得到两个词列表，对两个词列表进行合并去重，得到输入样本的所有词，选取词频作为特征值。最后进行向量化计算余弦值[1]。

3.2 Jaccard相似度

Jaccard 相似系数（Jaccard similarity coefficient）用于比较有限样本集之间的相似性与差异性。当 Jaccard 系数的数值越大，说明样本相似度越高。另外一种说法是用 Jaccard 距离表示相似度即：1−J(A，B)1—J(A，B)1−J(A，B)，但其本质一样，Jaccard 距离越大相似度越小。Jaccard 系数反映了两个向量间的关系[2]。Jaccard 系数很适合用来分析多个维度间的相似性，也多被用于推荐系统中用来给用户推荐相似的产品或业务。虽然 Jaccard 主要是在维度分析这样的稀疏向量中作用比较大，但是在文本相似度计算时也可用 Jaccard，诸如某些过滤相似度很高的新闻、网页去重、考试防作弊系统或者论文查重系统。拿到文本后首先进行分词，后进行求交集与并集，然后再进行做除法运算，得到 Jaccard 系数，而 Jaccard 距离为 1-Jaccard 系数。

$$J(A,B) = \frac{|A\cap B|}{|A\cup B|} = \frac{|A\cap B|}{|A|+|B|-|A\cap B|} \quad (3\text{-}2)$$

3.3 最小哈希相似度

集合上的最小哈希函数是基于全集的排序转换来定义，给定任意一个排列转换，集合的最小哈希值为在排列转换次序下出现的第一个集合元素。min hash 算法 LSH（Locality Sensitive Hashing，局部敏感哈希）中的一种，主要用来判断两个集合之间的相似性，因为这种方式计算的时间复杂度较低，所以适用于数据量大的集合。主要功能有两个：一是估算两个集合的相似度，min hash 可以控制计算值与实际值的误差范围，保证计算值的准确率；二是缩短计算的时间，传统的方法需要逐行逐项比对集合中的值，当数据量很大时，会造成运算时间呈几何倍数上涨，该算法中通过先提取出相似对，再进行逐项比对的方法，减少很多工作量，从而缩短了时间[3]。

3.4 SimHash相似度

Simhash 是一种指纹识别技术，它的特点是几乎重复的指纹在少量位元位置上是不同的。在文本相似度处理上，它可以提取出两个相似的文本近似的 hash 值。通过比较 hash 值得到两份文本的相似度。并且 LSH 的实现方式有多种，常用的就是 SimHash 算法。具体处理流程为：首先进行分词本系统采用 jieba 分词，求取哈希值通过 hash 算法把每个词变成 hash 值，通过 hash 生成结果，需要按照词的权重形成加权数字串，再把上面各个词算出来的序列值累加，变成只有一个序列串，最后进行降维。

3.5 TF-IDF相似度

TF-IDF（词频—逆向文件频率）技术一般用于信息检索与文本挖掘。TF-IDF 算法其本身是一种统计方法，用以评估一组词对于一个文件集或一个语料库中的其中一份文件的重要程度。主要是指单词的重要性与在文档中出现的次数成正比，但与此同时在语料库中出现的频率则成反比。TF-IDF 的主要思想是：TF-IDF 有两个指标，第一个是指词在文本中出现的频率，如果该词在文本反复出现，出现的次数越多，这个词就越重要。第二个是指该词在所有文本中出现的频率都很高，比如“我们”这个词。但是这个词显然就没那么地重要。TF-IDF 就是从这两个指标出发来判断该词在本文中的重要性。用这个方法计算出来的值就是 TF-IDF 值，TF-IDF 值与 DF 值成反比。转换为数学公式：TF-IDF =TF（词频）*IDF（逆文档频率）[4]。

使用 TF-IDF 算法求取文章相似度的具体流程如下：首先分别对两篇文章或文本进行分词，然后计算两篇文章或短文的 TF-IDF 值，再对其求相似度，如果相似度值越大就表示越相似。

4 系统架构与模块设计

4.1 系统总体架构

智能合同对比分析系统基于 Sentence-BERT 网络结构进行构建，具体架构如图 1 所示。

图 1 系统架构图

4.2 功能模块设计

4.2.1 合同类对比分析

该模块主要是通过上传文档进行对比分析，在合同审核流程，快速找出不同版本合同修改区域与版本差异，在合同盖章归档场景，有效识别实际签署纸质合同和电子版合同差异。

本模块处理流程如图 2 所示：

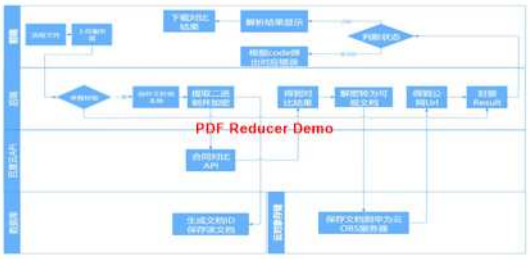

图 2 合同对比分析流程

4.2.2 其他类文本相似度检测

在自然语言处理的任务中，文本相似度是非常有用的工具，它可以解决很多实际问题，在诸如搜索引擎、推荐系统、论文鉴定、自动应答等领域都有着广泛的应用，本系统实现多种文本相似度算法可以供非计算机人士直接使用，快捷方便。

本模块处理流程如图 3 所示：

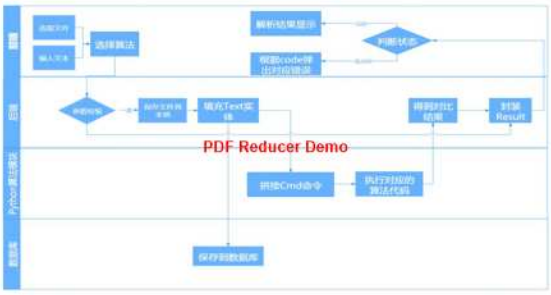

图 3 文本相似度流程

4.2.3 文本智能分析

文本智能分析模块主要是对文本进行简单的语言错误矫正，生成文本的词云，提取文章概要，提取文章的关键字。支持自行输入文本，上传的各类文档，并且支持移动端 OCR 识别文字进行分析。

本模块处理流程如图 4 所示：

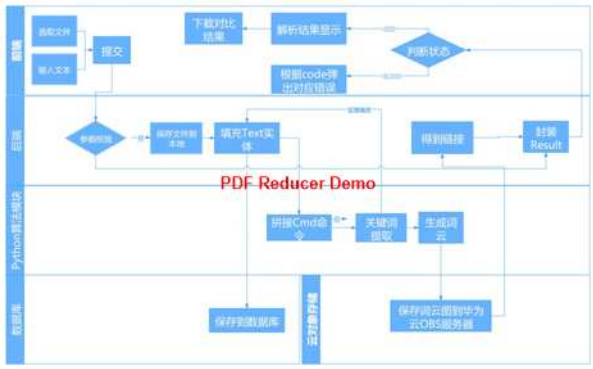

图 4 文本分析流程

5 结语

智能合同对比分析系统在企业管理和现代化、自动化办公中有着其独特的优点，可以帮助客户提高在进行文档相关工作中的工作效率，该系统的设计加入了前沿技术 BERT 模型进行语义相似性比较，这使得文档对比更加准确，更加合理，使用 Sentence-BERT 模型来做句子相似度匹配使得查找速度成倍减少，同时使用当前最为广泛使用的 spring boot 以及 vue 以前后端分离模式进行开发，提高开发效率。其系统功能设计基本符合实际需求，能够完成基本的文档辅助功能，如：合同和标书的对比分析，短文本相似度的求取以及文本的智能分析。

作者简介：罗通（1987.2—），男，海南乐东人，硕士，实验师，研究方向：大数据处理，人工智能。

【参考文献】

[1] 朱浩，连德富，左志宏，等.余弦相似度在高校综合信息系统中的应用[J].东南大学学报（自然科学版），2017，47（S1）：123-128.

[2] 夏竹青，王竹婷.基于余弦相似度的电子版实验报告管理系统[J].电脑知识与技术，2019，15（7）：100-102.

[3] 赵志靖，江荻.基于编辑距离的语言分类研究[J].语言研究，2020，40（2）：43-50.

[4] 宋颖毅，叶东升，王坤龙，等.无监督的问句相似度匹配方法[J].计算机应用研究，2020，37（S2）：69-72.