融合Synonyms词库的专利语义相似度计算研究

佟昕瑀[1]，廖佳伦[1]，路永和[2,*]
[1]中山大学信息管理学院, 广州, 中国, 510006
[2]中山大学人工智能学院, 珠海, 中国, 519082

摘要

一直以来，专利相似度计算和比较等工作都由专利审查员人工进行并做出准确判断。然而，以人工方式分析和研判专利的原创性、实用性以及是否侵权等工作需要投入大量的人力物力资源且效率较低。基于此，本文将ALBERT预训练模型用于专利的文本表示，并通过引入Synonyms近义词库增强专利文本的语义表达能力，探索一种基于语义知识库和深度学习的专利文本表示模型与相似度计算方法。实验结果表明，加入Synonyms近义词库消歧后的专利文本相似性度量的实验准确率有一定的提升。

关键词： 语义相似性；文本表示；自然语言处理；预训练模型

Patent Semantic Similarity Calculation by Fusing Synonyms Database

Xinyu Tong[1], Jialun Liao[1], Yonghe Lu[2,*]
[1]School of Information Management, Sun Yat-sen University, Guangzhou, China
[2]School of Artificial Intelligence, Sun Yat-sen University, Zhuhai, China

Abstract

Traditionally, patent examiners have relied on manual analysis to assess patent originality, practicality, and infringement, among other tasks related to patent similarity measurement and comparison. However, such analysis and judgment require significant human and material resources and low efficiency. Therefore, this paper utilizes the ALBERT pre-training model for text representation in patents and enhances the semantic expression capability of patent texts by introducing a Synonyms synonym library. It explores a patent text representation model and similarity calculation method based on semantic knowledge bases and deep learning. Experimental results demonstrate that incorporating Synonyms synonym library for disambiguation improves the accuracy of measuring patent text similarity.

Keywords: Semantic similarity , Text representation , Natural language processing , Pre-trained models

*路永和（通讯作者）：luyonghe@mail.sysu.edu.cn
本文系广东省重点领域研发计划项目“基于大数据智能的多层次知识检索关键技术研究及应用”（项目编号：2021B0101420004）的研究成果之一。

1 引言

专利作为科技文献的主要表现形式之一，从专利文本中可以获取当前创新技术，通过对专利文献进行分析，可以从中获得先进技术的细节，并进一步预测未来的技术发展趋势，以帮助相关群体进行关键的投资决策(Zhang et al., 2015)。在2021年的世界五大知识产权局（欧洲、日本、韩国、中国和美国）统计报告中指出，专利是用于保护发明创造的法律文件，专利数量被认为是衡量创新活动的重要指标，与日俱增的专利文献数量也导致了审查周期的延长(European Patent Office, 2023)。因此，各国专利局不断优化专利申请和审查程序以进一步缩短专利审查周期，从而达到鼓励创新、推动技术发展以及优化技术收益的目的。目前，专利授权以及专利侵权判定等大部分都是由专利审查员人工进行的，开展该项工作一方面需要具有不同领域背景的大量专利审查员，耗费较多的人力资源，另一方面人工判定耗时长且容易出现差错准确度难以保证(Liang and Tan, 2007)。

与其他文本相比较，专利文本是对某领域相关问题的创新深入研究，专业性强且会涉及大量领域词汇(Lee and Jeong, 2008)。专利文本的重点在于新技术的提出，因此，在技术的介绍部分往往会包含较多领域基础内容，在利用计算机判别时需要在文本的特征表达部分达到较高标准(Wen et al., 2021)。基于此，在判断专利文本的相似性时需要对领域词汇、生僻词和近义词进一步处理以提高文本的可读性。将知识库融入到专利文本相似性计算模型中，可以从文本语义理解的角度对专利文本进行更为准确的相似性计算。本文提出了一种基于Synonyms近义词库和ALBERT预训练模型的专利文本相似度计算方法，利用同义词库对专利文本的多义词、生僻词等进行解释或替换，使得文本成为更易于计算机神经网络理解的纯净文本，然后输入到预训练模型中，在下游任务中判断专利文本之间的相似度。

2 国内外研究现状

2.1 基于文本的专利相似度计算

基于文本内容的相似度计算方法中，最常用的方法是VSM模型，已有较多研究应用该方法或对其进行改进来计算专利文本的相似度。常用的方法是基于文本挖掘技术并结合专利关键词分析，计算专利文本元素的VSM加权相似度(Peng and Tan, 2010; Zhang et al., 2016; Arts et al., 2018)。

同样，Word2Vec和Doc2Vec模型也经常被用来表示专利文本，从而计算其相似度。Lee等人(2020)使用Word2Vec模型对特定产品领域的专利文本进行建模，并提出了一种产品潜在技术机会的分析方法。Xu等人(2018)用Word2Vec对专利文本进行训练和建模，并使用训练后的向量来计算专利领域词的相似度。Zhang等人(2018)以中文专利权利要求书为训练语料，用Doc2Vec模型计算专利权利要求文本的相似度，然后确定专利间的相似度。Cao等人(2018)使用Doc2Vec模型计算专利摘要相似度，并以此作为专利相似度参考。

在传统模型中加入语义信息可以有效提高模型计算专利技术相似度的准确率。具体方法包括设置不同的权重值以反映其他位置上的词的语义信息差异(Arts et al., 2021; Xia et al., 2018)，基于共享近邻聚类算法计算专利词的相似度(Jiang et al., 2016)，以及通过连接知识图谱中的专利术语和语义关系来优化专利相似度计算模型(Wang and Liu, 2022; An et al., 2021; Li et al., 2020)。通过融合专利文本特征的词位权重和领域相关性权重，并结合词向量加权方法和VSM文本表示法，将语义信息纳入专利相似度的测量中(Yu et al., 2019)。

在基于神经网络的专利相似度计算研究中，该研究将深度学习算法和专利模型树结合起来进行专利相似度计算，并采用连体LSTM算法(Mueller and Thyagarajan, 2016)进行专利应用，取得了良好的结果(Ma et al., 2018)。一些研究通过结合自然语言处理嵌入技术和最近邻相似性方法来衡量专利之间的技术相似性，将整个专利领域表现为一个技术网络(Hain et al., 2022)。基于文本的专利相似性计算方法已经发展得比较成熟，而结合深度学习的计算方法也不断提高了专利相似性计算模型的语义计算能力。

2.2 基于本体的专利相似度计算

基于本体的文本相似度计算方法的核心是基于人工构建的语义词典进行计算。目前，国际上使用最为广泛的语义词典是WordNet(Miller, 1995)，其包含了英文词汇与其实际意义的大量关系，展现了词与词之间的强大互联。中文词典方面，由梅家驹等人于1983年提出的《同义词林》(梅家驹, 1996)，以及其由哈工大进一步完善的扩展版本，是现在较为常用的语义

词典，其中包含了大量的中文词汇关系以及知识内容。此外，由董振东等人推出的HowNet（《知网》）(Dong and Dong, 2003)也是现在中文领域使用较多的语义词典。Synonyms近义词库的构建参照了《同义词词林》以及HowNet（知网）中所收录的词汇以及词汇之间的关系，但Synonyms相较于其他词典来说拥有更大的词汇量，对中文词义的解释更全面，并且更使用时更为方便。

刘影等(2011)通过对前人基于HowNet提出一种义原相似度的改进算法，实验结果表明，基于HowNet的方法可以和基于WordNet的方法取得一致的精确度；张思琪等(2017)基于WordNet本体，综合了大量前人的研究，改进了信息量IC计算模型，进而提出了两种混合式的语义相似度的计算方法，实验结果显示优于其他方法。Lu等(2023)在科技论文文本表示阶段融入WordNet词典，并以此作为后续引文推荐模型的输入，实验结果表明，融入WordNet的文本表示模型可以表达更多的语义信息，并且对后续模型输出有较好的提升效果。

现有研究证实，在文本表示阶段融入语义词典可以帮助模型学习到更多的语义信息，并且对多义词消歧有一定的帮助。基于此，本文将ALBERT预训练模型用于专利的文本表示，并通过引入Synonyms近义词库增强专利文本的语义表达能力，探索一种基于语义知识库和深度学习的专利文本表示模型与相似度计算方法。

3 专利文本表示及相似度计算模型构建

3.1 模型总体架构

本文构建的专利相似度研究模型由三个部分组成，分别是基于Synonyms的多义词消歧、基于ALBERT的文本表示和基于Softmax的相似性判断。模型结构如图1所示。

3.2 Synonyms近义词库

Synonym近义词库是一个开放在GitHub上的一个中文近义词工具包，它可以十分简便快捷的用于多种自然语言处理任务，并且对于不同领域的文本都具有很强的适应性，是一个通用型的近义词库。该工具包目前的词汇量已经达到了435,729个，包含了丰富的词语语义信息，主要可以根据其丰富的词汇量以及所包含的语义关系进行近义词检索等操作(Hai Liang Wang, 2017)。Synonym近义词库自发布以来进行了18次更新，最近一次更新时间为2022年5月5日。

随着英文语义词典高速发展，不断涌现了许多高质量词库，并且已有大量研究者将词典用于所研究的自然语言处理任务并对其进行泛化，加快了文本消歧、文本处理领域的发展，而纵观国内，相关的研究还偏少。因此，Synonyms近义词库的作者基于word2vec词向量训练了一个质量高、关系齐全的同义词库，将词语的表达规范化处理。在经典的信息检索系统中对于文档的检索都是基于严格的文字匹配操作得到的，查询算式与所匹配的文本之间字符必须是一一对应上的。而基于word2vec对大量的文本数据进行训练，可以获取到文本的上下文信息，获取上下文句子/词之间的语义关系，将词汇映射到低维的向量空间当中。因此，词汇之间的关系可以用词汇在向量空间中的距离来表示，那么度量相似性时也可以通过距离来度量。在Synonyms近义词库的构建中参照了《同义词词林》以及HowNet（知网）中所收录的词汇以及词汇之间的关系，但Synonyms相较于其他词典来说拥有更大的词汇量，对中文词义的解释更全面，并且更使用时更为方便(Hui et al., 2019)。图2是Synonyms近义词库运行时的实例，其中[$nearby_words$]是输入词的若干个近义词，以list的形式存储，并且按照与输入词的距离长度由近及远进行排列；[$nearby_words_score$]是每个近义词所对应的距离得分，该得分的区间为（0-1），越接近于1则说明与输入词越为接近；[$size$]为返回的词数量，在Synonyms中该值默认为10。

3.3 ALBERT预训练模型

自从BERT问世以来，越来越多的研究将研究目光转向了预训练模型，并且为了使得模型的训练效果更好，许多研究人员选择在模型中加大参数量以取得更好的模型表现。然而，如果无限制的去增加参数量则一定会带来一些附加问题，如参数量越大则模型训练的速度就越慢，对计算机性能的要求也越高，当模型的参数增加到一个临界点时不仅不能继续提升模型的效果，甚至会适得其反的因为参数的过大导致模型效果变差(Lan et al., 2020)。为解决模型参数量过大、内存占用过多等问题，ALBERT团队提出了因式分解嵌入层矩阵和跨层参数共享两种能够大幅减少预训练模型参数量的方法，此外还提出用Sentence-order prediction（SOP）任务代

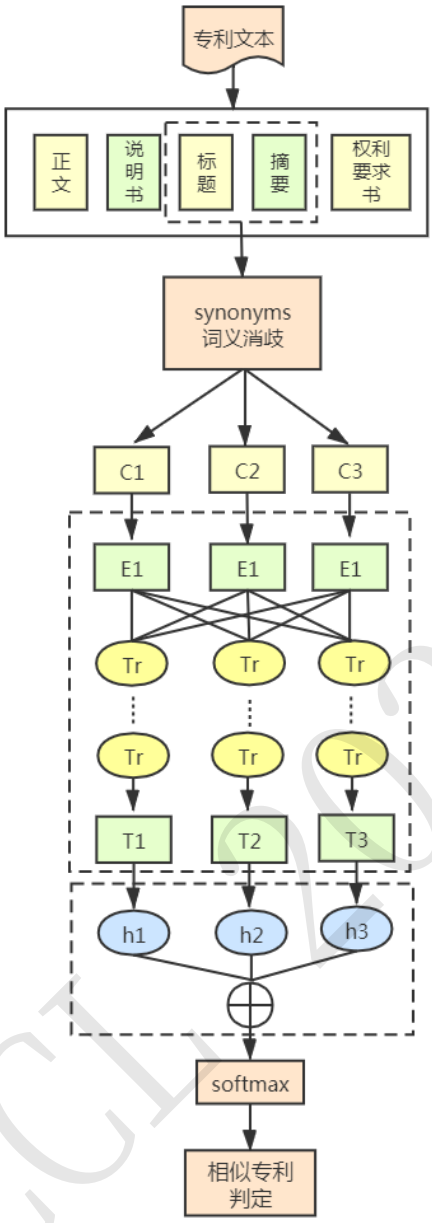

图 1: 基于Synonyms和ALBERT的专利相似度研究模型

```
Synonyms.nearby（人脸，10）=（
    [“图片”,“图像”,“通过观察”,“数字图像”,“几何图形”,“脸部”,“图象”,“放
     大镜”,“面孔”,“Mii”],
    [0.597284, 0.580373, 0.568486, 0.535674, 0.531835, 0.530095, 0.525344, 0.524009,
     0.523101, 0.516046]）
```

图 2: Synonyms近义词库实例

第二十二届中国计算语言学大会论文集，第341页-第351页，哈尔滨，中国，2023年8月3日至5日。

替BERT中的Next-sentence prediction（NSP）任务，经实验证实其在许多自然语言理解的任务和应用中取得了客观的结果。

因式分解嵌入层矩阵顾名思义可知该矩阵是对嵌入层的输入词向量做因式分解。词嵌入训练不仅仅只局限于某一部分的语境，而是将眼光放至了文本全局，使得最后生成的向量能够从本质上代表该词的全局语境。BERT 中的词嵌入层大小E和隐藏层的大小H相等，其隐藏层学习得到的是上下文相关向量。而在BERT的预训练模型中，可以将嵌入层大小与隐藏层大小根据实际任务要求和效果进行优化，比如将H设置的较大，则可以包含更多的信息，甚至可以根据需要设置为$H >> E$。此外，在自然语言处理任务中字典大小V通常达到上万，随着H的增大，而如果EaH则会导致词嵌入矩阵的无限扩大。为解决以上问题，ALBERT在BERT的基础上予以改进，通过引入数学因式分解的方式，把E和H分开设定，这样把原本较大的矩阵分解成了两个小矩阵，参数量有十分明显的减少，当$H >> E$的时候，参数削减更加明显。例如：$V = 30000$，$E = 128$，$H = 768$，则原参数量$V * H = 30000 * 768 = 23,040,000$，削减后$V * E + E * H = 30000 * 128 + 128 * 768 = 3,938,304$，参数变成了原来的1/6。

深度学习领域的重要参数压缩方式就是参数共享，其在很多深度学习神经网络中都有应用，不同的神经网络模型会根据自己实际的模型结构以及实现效果在不同的结构位置实现参数共享。而ALBERT是对Encoder中每一层之间参数共享，即多个层使用相同的参数，默认将全部的参数都进行共享，而不仅仅局限于前馈层或者注意力层。虽然该方式的性能相比于其他两种降低了1个百分点，但在参数量的减少上有着显著的效果。

在BERT的上下文判定中，不仅会考虑到两个句子之间的连贯性，还会考虑到两个句子的话题。BERT中所包含的NSP任务将注意力放在句子所表征的话题是否一致上，因此模型的判断是否为上下文任务变为了一个非常简单的主题匹配，而未考虑句子之间是否连贯，这使得NSP的任务无法学习更多的文本句子信息。因此，ALBERT中提出了SOP来取代NSP，具体来说，SOP分为正反两例，正序的判断方式与NSP一样，而反例则是将原本相邻的两个句子交换顺序，这样模型可以对句子的连贯性做出更多的吸收。在文本的实验中，ALBERT预训练模型充当着文本语义表示的角色，将专利文本输入至ALBERT模型后能够获得与其他模型相比更为丰富的语义表达信息。同时，也规避了大规模的参数导致模型占用大量计算资源。当使用ALBERT预训练模型判断两个句子在语义上是否相同时，可利用标注样本对预训练模型的参数进行微调以达到更好的效果，即通过持续的预训练不断学习数据集特征，优化下游任务结果。

4 实验及结果讨论

4.1 数据采集与预处理

4.1.1 数据采集

本文所爬取数据集来源于Google Patent上通信技术领域和文本处理领域的专利全文数据，所使用的编程语言是Python，并安装了selenium包，使用队列技术获取专利全文数据。在爬取过程中主要参考了专利审查员对于专利的标记识别，专利审查员标记的引用专利在技术上有较强的相似性(Chen, 2017; Lu et al., 2020)。除了专利审查员人工审查的内容之外，本文为扩展数据内容，将那些并没有被专利审查员判定为相似的专利对而Google自动判定为相似的专利对也纳入到了本文的数据获取队列中。当爬取的专利数量超过设定的阈值或者人工检测发现专利所涉及的主题与本文所确定的两个领域之间关联度很小的时候则立即停止对专利文本的爬取。本文一共爬取了通信技术和文本处理的两个专业领域共2620篇专利数据，构建实验数据集用于进行专利文本表示及相似度计算实验。表1是实验数据集相关字段含义描述。 由于本文的下游任务为专利相似性的判断，其为一个二元分类任务，只需要判断两个专利文本之间相似与否，因此在数据集的采集上获取了Similar字段，该字段所表示的含义为与该专利在内容上为相似专利的专利文本，该相似性判断由Google Patent所设置的程序和专利审查员共同进行判断，其中所判断的相似专利是与该专利不含有引用关系的专利文本。

4.1.2 数据预处理

文本预处理是后续进行文本表示的基础，预处理的效果会在很大程度上影响到后续相关任务的有效性和准确性。本文数据预处理中最重要的步骤是文本分词，为完成这一步，常用的方式时使用分词工具将中文文本分为最小字词，然后将其转换为词表中对应的数字id。为了取得

表 1: 实验数据集字段描述

字段名	字段含义
Patent_no	专利公开号，是表示专利唯一性的标识符之一。
Title	专利的标题。
Abstract	专利的摘要。
Citations	专利引用的专利列表，特指专利审查员的引用。
Cited	被引专利列表，即引用该专利的专利列表。
Similar	与该专利相似的专利列表，由Google Patent自动识别，不包含与该专利存在引用关系的专利。

更好的分词效果，本文首先使用NLPIR分词软件提取专利文本中的关键词，并人工筛选使用频率高、实际意义强的关键词加入用户字典辅助后续的分词过程。分词阶段采用了目前使用范围最广泛且操作方式简单的jieba分词系统。分词之后的文本数据中仍然包含了一些停用词和标点符号，这些对于我们的文本分析任务来说属于杂质，应当导入停用词表对其进行剔除。

为了便于后期的实验过程，针对本文的研究内容对专利文本进行具体处理。首先将每一个专利的编号、标题、摘要以及它相似文本的专利编号提取出来，并做简单的预处理工作。由于专利文本的标题通常较短，因此可以直接使用。而摘要部分句子比较长，但由于摘要的前半部分通常是包含了文本主题的主要信息，并且ALBERT模型的句子接受长度最多为512个字符，因此截选了摘要中的前512个字符输入至神经网络模型中。

4.2 实验流程及参数设置

本文进行专利文本表示和相似度研究构建了8组对比实验，包含了基于深度学习的神经网络模型及其对比模型。分别通过在专利文本表示及其相似度计算的各个环节采取不同的方法再进行组合从而得到本文所选择的对比实验模型。最终所得到的实验模型可归纳为两组，第一组实验为只使用ALBERT预训练模型和Softmax分类器对专利文本进行文本表示和相似度计算与使用传统向量空间模型进行文本表示和余弦相似度计算进行对比实验，第二组实验为基于Synonyms词库和ALBERT预训练模型以及Softmax分类器与上述传统模型进行比较实验。其中，将余弦相似度判断的阈值设定为0.8，若两篇文本余弦值大于0.8则判定其为相似文本，反之则不相似。每一组实验中专利文本的呈现方式都分为“标题”、“摘要”和“标题+摘要”三种。具体的对比实验流程图如下图3所示。

4.2.1 Synonyms文本消歧实验

文本消歧实验通过计算文本单词的tf-idf值加以人工筛选判断得到专利文本的前5个关键词，将关键词输入到Synonyms词典中并将其近义词追加到关键词后，由于有些词会有不止一个关键词，为了便于操作，本文选取synonyms中给定排序中的第一个近义词。筛选完成后将近义词添加到专业词汇后面。之所以采用直接添加的方式，是因为如果直接将原词替换成为其同义词可能会影响到原词在文本中所表达的意思使其不够准确。而直接在后面添加，该同义词与原词具有同样的意思表达，可以将原词的意思更为扩展便于计算机理解，并且不会影响到上下文的原意，保持句子的整体通顺。

4.2.2 ALBERT预训练模型构建

模型中的参数来源主要为在训练过程中自动获取，并且根据实际的模型运行情况结合人工经验对参数进行了必要的调整。表2列举了本文模型涉及的主要超参数类型及其描述。通过对已有ALBERT预训练模型的相关研究进行调研，以及在实验过程中反复测试调参，最终发现在本文的研究框架和内容下，每批训练数据设置为6时，可以与实验的输入长度512有较好的契合程度，并且对于数据集的遍历次数也可以达到适中的程度，模型能够达到一个较快的迭代收敛速度，同时对于计算机的内存消耗不大。实验设置迭代轮数为4，最大序列长度为512，在模型的优化器使用上选择了基于低阶矩阵估计的Adam，其实现过程限制较少，对其运行环境的性能要求不严苛，较为适合后续实验的调整和迭代。同时该算法还有利于训练的准确性和稳定性，防止其他干扰的杂质影响。

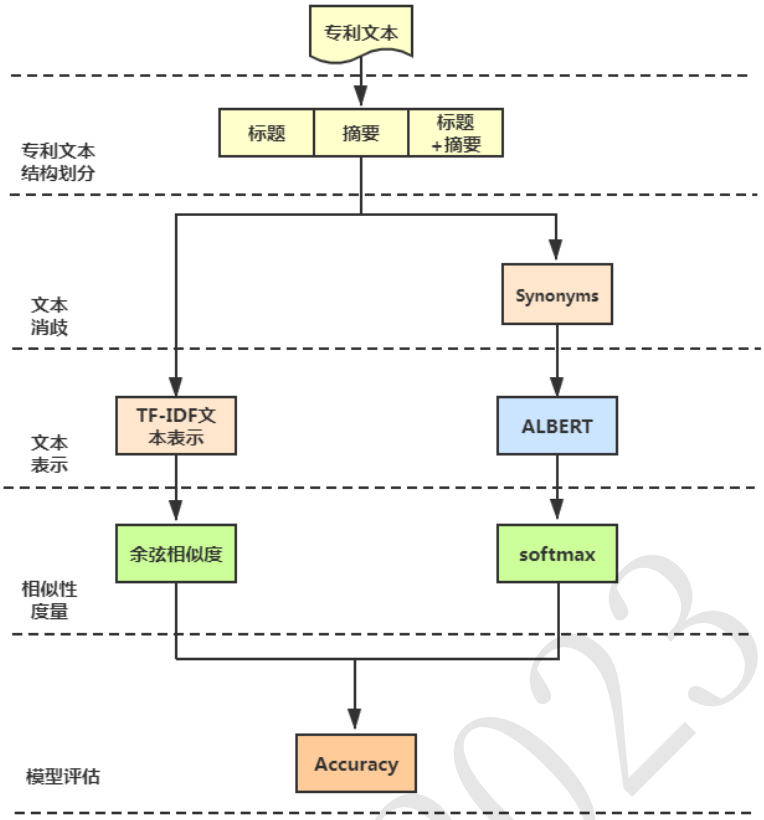

图 3: 实验模型及其对比模型

表 2: 实验参数

参数名称	参数设置	说明
Size	32000	所用词表大小
Embedding size	128	词嵌入层的大小
Hidden layer num	10	输入层的层数
Hidden size	768	输入层和池化层的大小
Hidden gro	1	Encoder的分割组数
Mul attention heads	12	多头自注意力层头数
Fully connected layer	3072	Encoder全连接层大小
Scaling	1	缩放比例
Activation function	gelu	激活函数
Dropout probability	0	全连接层dropout概率
Attention dropout probability	0	注意力层dropout概率
Max embeddings	512	最大长度
Vocab size	21128	类别数量
Train:Dev:Test	6:2:2	训练集、验证集和测试集比例

为防止过拟合的现象发生，本文在ALBERT实验代码编写过程中进行了一些调整以取得更好的实验效果。加入BatchNormld函数用以防止梯度消失，避免出现权重的过大或过小等情况；每训练50个import就会让学习率降低一定的参数值；虽然ALBERT模型本身为了减少对计算机内容的占用因此进行了移除了dropout的操作，但是实验中发现加入dropout函数对于防止过拟合存在着较好的效果，因此在本文实验过程中保留了dropout。

本文的实验所采用的系统为$Windows10_64$ 位，硬件配置为$Interi7-9700k$处理器、8GB内存，显卡型号为七彩虹$iGameGeForceRTX2080TiAdvance$。本文所有的实验所使用的编程语言均为Python，编写代码所使用的是Pycharm。其中神经网络模型部分由pytorch库中的函数搭建。

4.3 实验结果及分析

本文选择准确率accuracy来评价模型在专利文本相似度计算方面的效果。表3是具体的实验结果数据，共设置了两组对比实验，分别是：（1）ALBERT与传统的向量空间模型与余弦相似度组合算法的比较；（2）ALBERT与加入Synonyms近义词库的ALBERT模型对比。同时也比较了专利标题、摘要与标题+摘要对专利相似度分类实验的影响。

表 3: 实验结果

模型	准确率
title+tf-idf+cos	0.474763
abstract+tf-idf+cos	0.536277
title-abstract+tf-idf+cos	0.526813
title+ALBERT	0.824921
abstract+ALBERT	0.837539
title-abstract+ALBERT	0.850157
synonyms+ title+ALBERT	0.837838
synonyms+abstract+ALBERT	0.851351
synonyms+title-abstract+ALBERT	0.864864

从文本组合方面来看，用摘要部分进行文本表示和相似度匹配的准确率要比单用标题高，这说明专利文本的摘要部分包含了较多的主题信息，能够较为全面的反应出专利所表达的主题内容。而采用标题+摘要的方式所取得的效果则会更佳，这可能是因为标题和摘要之间互相补充互相配合更好的体现出了专利文本所描述的主题细节，同时两者的结合也能避免一个词存在多种意义而使得模型判断失误的情况出现。

通过第一组对比实验可以得出，将专利文本输入到ALBERT深度学习模型中进行文本表示以及相似度判断的准确率要远远高于使用传统的方式，具有30%左右的准确率提升，作为一个二分类的下游任务，ALBERT在实验采用的数据集上取得较好效果。究其原因可能在于专利文本具有较强的专业性和语言理解性，其文本中出现的内容通常为某一领域的新技术、新方法。而tf-idf文本表示方式是基于词频统计的模型，无法反映词与词、句子与句子之间的语义关系，难以有效捕捉到文本的主题信息。此外，余弦相似度的计算也只是基于向量进行简单的数学公式运算，其在计算过程中并不能考虑到相关的语义信息，导致使用tf-idf+余弦相似度来计算专利文本相似度的准确率较低模型效果较差。而使用ALBERT模型进行专利文本表示和相似度研究时，其能够通过自注意力机制关注到上下文的文本信息从而预测当前单词，可有效对文本进行表示。

从加入Synonyms近义词库消歧的结果来看，在标题、摘要、标题-摘要的三种文本组合与ALBERT预训练模型的实验中，加入了Synonyms近义词库做词义消歧的效果相比于未加入时准确率提升了1%左右，说明在加入外部知识库进行语义消歧确实能够对ALBERT预训练模型进行文本表示和下游文本相似度计算任务有一定的提升。但是从实验结果中看到，加入知识库后的提升并未达到最为理想的效果，可能有以下两点原因：（1）同义词替换与添加的方式可能会影响到句子原始语义的表达，甚至会在某些情况下反而影响到句子语言的通顺程度，不利于ALBERT预训练模型对文本特征的提取，因此对于ALBERT模型来说帮助并不大；（2）在

利用synonyms进行同义词替换和添加时并没有考虑到上下文语境，将导致多元的信息变得单一化，可能进一步模糊了句子间的边界。由于在知识库的使用上未能将其更加融入进神经网络模型的训练过程中，会在一定程度上弱化知识库的使用效果，从而导致整体的提升不大。本研究所得到的结果也可证明采用这种方式引入外部知识库仍然能够在文本语义的消歧和扩展上起到提升的效果。

5 总结

为了在大量的专利文本中找到其相似专利，便于研究者在开展研究前查询资料以及帮助专利审查员判定专利是否侵权，本文提出了一种结合语义信息和深度学习共同进行文本表示和相似度计算的方法，并验证了该方法在专利文本上应用的可行性和有效性。

本文选取了目前模型效果和性格各方面综合较为优秀的ALBERT预训练模型用于专利文本表示和相似度计算任务领域，通过ALBERT模型对专利文本特征的提取得到更为准确的专利文本表示，并连接Softmax分类器对下游的专利文本相似度计算任务做出更为精准的判断。在此基础上，引入开源的Synonyms近义词库在专利文本输入至ALBERT预训练模型之前对专利文本进行消歧以及语义扩展。最终通过采集谷歌专利库中文本处理和通信技术两个领域的专利文献验证实验效果，发现在使用Synonyms近义词库消歧后再将专利文本输入到ALBERT预训练模型中的实验效果比起只使用ALBERT预训练提升了1%，即对专利文本语义表达能力进行了一定程度的增强。

此外，实验部分根据专利文献的行文结构特点，截取了专利文本的标题和摘要两个部分进行文本组合，最终得到“专利标题”、“专利摘要”、“专利标题+摘要”三种文本组合，将其分别输入至实验模型中可以发现“专利标题+摘要”的效果要略好于其他两种组合方式，说明输入模型的文本长度可以影响模型特征提取效果。

综上所述，在专利文本表示及相似度计算领域引入预训练模型能够取得较好的实验效果，同时引入外部的知识库对文本的语义信息进行增强也能够对预训练模型起到一定的辅助作用，提升模型的效果。在未来对专利原创性、新颖性以及是否侵权等做出判断时，则可以将大部分工作任务转向由计算机程序自动处理，而减少对于人力资源的投入。

参考文献

Xin An, Jinghong Li, Shuo Xu, Liang Chen, and Wei Sun. 2021. An improved patent similarity measurement based on entities and semantic relations. *Journal of Informetrics*, 15(2):101135.

Sam Arts, Bruno Cassiman, and Juan Carlos Gomez. 2018. Text matching to measure patent similarity. *Strategic Management Journal*, 39(1):62–84.

Sam Arts, Jianan Hou, and Juan Carlos Gomez. 2021. Natural language processing to identify the creation and impact of new technologies in patent text: Code, data, and new measures. *Research Policy*, 50(2):104144.

Qi Cao, Wei Zhao, Yingjie Zhang, Shujun Zhao, and Liang Chen. 2018. Comparative study of patent documents similarity detection on deep learning of doc2vec based methods. *Library and Information Service*, 62(13):74–81, 1.

Lixin Chen. 2017. Do patent citations indicate knowledge linkage? the evidence from text similarities between patents and their citations. *Journal of Informetrics*, 11(1):63–79.

Zhendong Dong and Qiang Dong. 2003. Hownet - a hybrid language and knowledge resource. In *International Conference on Natural Language Processing and Knowledge Engineering, 2003. Proceedings. 2003*, pages 820–824.

European Patent Office. 2023. Ip5 statistics report 2021 edition. https://www.fiveipoffices.org/sites/default/files/2023-01/IP5%20Statistics%20Report%202021_1.pdf.

Hu Ying Xi Hai Liang Wang. 2017. 中文近义词工具包synonyms.

Daniel S. Hain, Roman Jurowetzki, Tobias Buchmann, and Patrick Wolf. 2022. A text-embedding-based approach to measuring patent-to-patent technological similarity. *Technological Forecasting and Social Change*, 177:121559.

Yueting Hui, Yijia Xia, Zihe Chen, and Xin Tong. 2019. Short text clustering algorithm based on synonyms, and k-means. *Computer Knowledge and Technology*, 15(1):5–6, 1.

Lixue Jiang, Duo Ji, and Dongfeng Cai. 2016. Measuring term similarity based on internal semantic role in patent text. *Journal of Chinese Information Processing*, 30(4):37–43, 1.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. Albert: A lite bert for self-supervised learning of language representations.

Bangrae Lee and Yong-Il Jeong. 2008. Mapping korea's national r&d domain of robot technology by using the co-word analysis. *Scientometrics*, 77:3–19.

Changyong Lee, Daeseong Jeon, Joon Mo Ahn, and Ohjin Kwon. 2020. Navigating a product landscape for technology opportunity analysis: A word2vec approach using an integrated patent-product database. *Technovation*, 96-97:102140.

Quanjia Li, Baoan Li, Xindong You, and XUeqiang Lyu. 2020. Computing similarity of patent terms based on knowledge graph. *Data Analysis and Knowledge Discovery*, 4(10):104–112, 10.

Yanhong Liang and Runhua Tan. 2007. A text-mining-based patent analysis in product innovative process. In Noel León-Rovira, editor, *Trends in Computer Aided Innovation*, pages 89–96, Boston, MA. Springer US.

Ying Liu, Li Chen, Zilin Song, Qingchao Dong, Xinghua Chen, Weixing Zhu, and Jixian He. 2011. An improved ontology-based method to measure similarity between concepts. *Journal of Nanjing University of Posts and Telecommunications(Natural Science)*, 31(6):60–66, 1.

Yonghe Lu, Xin Xiong, Weiting Zhang, Jiaxin Liu, and Ruijie Zhao. 2020. Research on classification and similarity of patent citation based on deep learning. *Scientometrics*, 123:813–839, 02.

Yonghe Lu, Meilu Yuan, Jiaxin Liu, and Minghong Chen. 2023. Research on semantic representation and citation recommendation of scientific papers with multiple semantics fusion. *Scientometrics*, 128:1367–1393, 01.

Chunyan Ma, Tong Zhao, and Hao Li. 2018. A method for calculating patent similarity using patent model tree based on neural network. In Jinchang Ren, Amir Hussain, Jiangbin Zheng, Cheng-Lin Liu, Bin Luo, Huimin Zhao, and Xinbo Zhao, editors, *Advances in Brain Inspired Cognitive Systems*, pages 633–643, Cham. Springer International Publishing.

George A. Miller. 1995. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41, nov.

Jonas Mueller and Aditya Thyagarajan. 2016. Siamese recurrent architectures for learning sentence similarity. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, page 2786–2792. AAAI Press.

Jidong Peng and Zongyin Tan. 2010. A text mining-based patent similarity measurement method and its application. *Information Studies:Theory Application*, (12):114–118, 1.

Zihong Wang and Y. Liu. 2022. Sea-ps: Semantic embedding with attention to measuring patent similarity by leveraging various text fields. *Journal of Information Science*.

Chaodong Wen, Cheng Zeng, Junwei Ren, and Yan Zhang. 2021. Patent text classification based on albert and bidirectional gated recurrent unit. *Journal of Computer Applications*, 41(2):407–412, 2.

Bin Xia, Baoan Li, and Xueqiang Lyu. 2018. Calculation of patent text similarity based on word location and semantic information. *Computer Engineering and Design*, 39(10):3087–3091, 1.

Kan Xu, Yuan Lin, Chen Qu, Bo Xu, and Hongfei Lin. 2018. Research on patent query expansion methods using word embedding. *Journal of Frontiers of Computer Science Technology*, 12(6):972–980, 1.

Yan Yu, Lei Chen, JInde Jiang, and Naixuan Zhao. 2019. Measuring patent similarity with word embedding and statistical features. *Data Analysis and Knowledge Discovery*, 3(9):53–59, 1.

Haichao Zhang and Liangwei Zhao. 2018. Judge chinese patents similarity based on doc2vec. *Technology Intelligence Engineering*, 4(2):64–72, 1.

Longhui Zhang, Lei Li, and Tao Li. 2015. Patent mining: A survey. *SIGKDD Explor.*, 16:1–19.

Yi Zhang, Lining Shang, Lu Huang, Alan L. Porter, Guangquan Zhang, Jie Lu, and Donghua Zhu. 2016. A hybrid similarity measure method for patent portfolio analysis. *Journal of Informetrics*, 10(4):1108–1130.

Siqi Zhang, Weiwei Xing, and Yuanyuan Cai. 2017. A wordnet-based hybrid semantic similarity measurement. *Computer Engineering and Science*, 39(5):971–977, 1.

梅家驹. 1996. 同义词词林. 上海辞书出版社.