

基于知识注入提示学习的专利短语相似度计算

邓远飞¹, 李加伟¹, 蒋运承^{1, 2}

(1. 华南师范大学计算机学院, 广东 广州 510631; 2. 华南师范大学人工智能学院, 广东 佛山 528225)

摘要: 专利是授予发明者在一定时期内保护其发明的法定权利, 在当今的社会活动中发挥着重要作用。然而现有研究并未针对专利相似度数据进行适配优化, 导致其应用在专利短语相似度匹配任务中效果不佳。已有研究表明, 在低资源的场景下, 提示学习将文本片段(模板)作为输入, 将分类问题转换为掩码语言建模问题, 其关键的一步是在标签空间和标签词空间之间构造一个投影。提出一种基于知识注入的提示学习方法, 将其应用于专利短语相似度匹配计算任务。为解决专利短语信息不足的问题, 利用专利短语中的相似度标签信息, 使用知识增强专利短语与标签信息。首先通过实体链接技术建立专利短语与外部知识的关联关系; 然后设计一种基于实体影响度的邻域信息过滤机制, 用于缓解专利短语信息不足的问题; 最后考虑不同外部知识对专利短语相似度计算的影响, 设计应用于专利短语的多种增强提示文本。实验结果表明, 该方法的 Pearson 相关系数(PCC)和 Spearman 相关系数(SRC)相较次优对比方法分别提升 6.8% 和 5.7%。

关键词: 专利短语; 相似度计算; 知识注入; 提示学习; 提示文本

中图分类号: TP391

文献标志码: A

DOI: 10.19678/j.issn.1000-3428.0067595

Similarity Computation of Patent Phrases Based on Knowledge Injection Prompt Learning

DENG Yuanfei¹, LI Jiawei¹, JIANG Yuncheng^{1, 2}

(1. School of Computer Science, South China Normal University, Guangzhou 510631, Guangdong, China;

2. School of Artificial Intelligence, South China Normal University, Foshan 528225, Guangdong, China)

【Abstract】 A patent is a legal right conferred to inventors to protect their inventions for a limited time, and it plays a crucial role in present-day social activities. Existing research has not optimized the adaptation of patent similarity data, which has negatively affected matching patent phrase similarity. Previous research has shown that in low-resource scenarios, prompt learning uses text fragments (i.e., templates) as input, transforming the classification problem into a mask language modeling problem; here, a key step is to construct a projection between the label space and label word space. This study presents a knowledge-based prompt learning method and applies it to the similarity matching of patent phrases. To solve the problem of insufficient information related to patent phrases, this study uses similarity label information in patent phrases and knowledge to enhance the patent phrases and label information. This study first establishes the relationship between patent phrases and external knowledge using entity-linking technology. The study then designs a neighborhood information filtering mechanism based on the degree of entity influence to expand the problem of insufficient patent phrase information. Finally, based on the effects of different types of external knowledge on the similarity calculation of patent phrases, the study generates a variety of enhanced prompt text applied to patent phrases. Experimental results show that the Pearson Correlation Coefficient (PCC) and Spearman Rank Correlation (SRC) of the proposed method are increased by 6.8% and 5.7%, respectively, as compared with the suboptimal method.

【Key words】 patent phrase; similarity computation; knowledge injection; prompt learning; prompt text

0 引言

专利是知识产权的一种形式, 对于管理信息和知识、研发活动以及减少研究周期和费用都有帮助^[1-2]。专利的申请数量每年都在持续增长, 需要高效的系统来管理如此庞大的数据。现有研究提供了

许多方法来解决这个问题, 如有效处理、分析、分类和存储专利数据的方法^[3-4]。

最近几十年来, 专利法和专利技术的创新实证研究受益于质量越来越高的专利数据集, 这些数据集在国家、企业、团队和个人层面都被广泛应用。专利数据有多种用途, 如作为知识输入的代理或专利

收稿日期: 2023-05-10 修回日期: 2023-07-04

基金项目: 国家自然科学基金(61772210, U1911201)。

通信作者 E-mail: dengyf@m.scnu.edu.cn

价值的度量,也被用作创新代理的度量。由于专利数据量庞大,因此需要大量的人力成本来研究和搜索专利。目前,专利管理和搜索的方式主要是基于传统的数据库,即技术人员需要识别目标的关键字并进行搜索,然后从数据库中检索专利^[5]。因此,为了有效地对专利文本进行语义相似度计算,需要设计一种准确的专利语义表示方法。

USPTO(United States Patent and Trademark Office)数据在专利分析、经济学、起诉和诉讼工具等领域频繁使用。大多研究中使用的专利分类系统是参照国际专利分类法(IPC)和联合专利分类(CPC)所建立的。目前,针对 USPTO 数据的专利相似度计算及检索相关研究,仅利用自然语言处理相关技术来实现^[6]。

随着预训练语言模型(PLM)的不断发展,BERT^[7]等具有动态语义表征能力的模板逐渐成为文本表示领域的通用模型,使得解决下游任务的策略从重新训练模型转变为根据任务调整其参数。然而,在预训练和微调的过程中,不同阶段的优化目标不同。于是,起源于生成式预训练 Transformer(GPT)^[8]并由模式探索训练(PET)^[9]等发展而来的“提示学习”^[10]新范式应运而生。值得注意的是,已有很多工作尝试将外部知识融入以 BERT 为代表的预训练语言模型中,但主要集中在常识领域知识^[11]和开放领域知识^[12]。清华大学曾通过外部知识图谱扩展标签映射的方法,即知识集成的提示调优(KPT)^[13],在关系抽取、文本分类等任务中取得了较大的性能提升。此外,文献[14]研究表明,并不是所有的外部知识都能带来增益。目前,该领域主要面临 2 个挑战:1)如何有效地植入外部知识,解决专利短语信息不足的问题? 2)如何充分利用专利短语中的标签信息,解决知识噪声和异构性问题? 如果这些技术被有效地应用于专利检索、专利分类等任务,专利推荐、自动专利质量预测等应用在未来甚至可以实现。

针对专利短语相似度语义匹配问题,本文首先提出基于知识注入的提示学习方法并应用到专利短语相似度计算中;其次引入外部知识图谱 Wikidata,提出基于实体影响度的邻域过滤机制,获取相关实体的邻域特征信息,通过外部信息源扩充标签与专利短语信息,提高分类准确度同时解决专利短语信息不足的问题;接着应用交叉熵损失函数来训练所提模型,将扩展标签词的分数的映射到标签分数上。此外,本文设计一种有效的提示生成文本,并验证其在现有数据集上的有效性。

1 相关工作

1.1 专利相关工作

近年来,许多机器学习技术已应用于专利相关研究,主要集中在专利检索、专利分类和相似度计算中,提高了专利分析工作的效率^[2]。文献[6]提出一种两阶段专利检索方法,通过考虑权利要求的结构,将检索到的文档重新排序,取前 N 个文档,这意味着权利要求在专利中具有重要意义。随后,文献[15]通过使用 USPTO 专利数据,利用一种引文分析方法改进专利检索技术。文献[16]提出一种基于聚类的专利检索技术,通过使用 IPC 代码提供额外的数据,满足用户的信息需求。

另一方面,尽管专利具有半结构化的性质,但是大多数现有的关于专利分类的研究工作都侧重于文本信息。如今,深度学习方法(如卷积神经网络)在图像处理、语音识别等领域取得了巨大进展,但尚未应用于专利分类任务。文献[17]提出了一种基于卷积神经网络和词向量嵌入的深度学习算法 DeepPatent 并用于专利分类。PatentBERT 专注于微调预先训练的 BERT 模型,该模型仅使用专利的第一项权利要求,并在 CPC 子类级别的 656 个标签上取得了显著的成果。Patent2Vec^[18]在专利分类工作中使用标签的文档学习专利的低维表示,从多视图的角度执行专利分类任务。

1.2 提示学习

经过微调的 PLM^[19]在各种自然语言处理任务中取得了巨大的成功。PLM 可学习关于语言的语法^[20]、语义^[21]和结构^[22]信息,在智能问答^[23]、文本分类^[24]和机器翻译^[25]等领域广泛应用。然而,PLM 仍然阻碍下游任务对预训练知识的充分利用。为此,受 GPT-3^[26]的启发,提示学习方法应运而生,其将下游任务转换为一些形式化目标并取得了优异的性能^[17]。

提示学习^[27-28]是最近出现的模型训练方法,最初由人工设计模板。文献[29]提出的 PET 借助自然语言构成的模板,使用 BERT 的掩码语言模型来进行预测。但是,PET 方法的局限性在于需要人工选择模板,而且 PET 的准确率严重依赖模板的质量。文献[30]提出前缀调优(Prefix-tuning),其放弃模板由自然语言构成这一常规要求,使用连续空间内的向量作为模板。GPT-3、ChatGPT^[26]相关研究表明,通过快提示调优和上下文学习,大规模语言模型可以在低数据状态下

实现优异的性能。P-tuning^[31]利用可训练的向量来自动构建提示文本模板,重点解决提示调优在小模型上效果不佳的问题。手工定义或自动搜索得到的标签词映射有主观性强、覆盖面小等缺点,KPT^[12]通过外部知识库扩展描述器的提示调优,用于提高文本分类的准确率。同时,KPT 也为在提示学习下引入外部知识提供了参考。

2 背景知识

2.1 专利短语相似度语义匹配数据

本文使用的专利短语相似度语义匹配数据集(PPSD)是一个人类专家评级的 CPC 上下文专利技术短语(术语)到短语匹配数据集。表 1 所示为专利短语相似度匹配数据集的一个示例,完整的数据集可以通过 Kaggle(<https://www.kaggle.com/datasets/google/google-patent-phrase-similarity-dataset>)公开获取。

表 1 专利短语相似度语义匹配数据示例

Table 1 Example of semantic matching data for patent phrases similarity

Anchor	Target	Context	Rating	Score
acid absorption	acid absorption	B08	exact	1.00
gasoline blend	petrol blend	C10	synonym	0.75
gasoline blend	gasoline mixture	C10	hypernym	0.50
faucet assembly	tap inputs	A22	not related	0.00

专利短语相似度语义匹配数据集包含近 50 000 个评级短语对,每个短语对都有一个 CPC 类作为上下文,分为训练集(75%)、验证集(5%)和测试集(20%)。数据集中存在成对的专利短语(Anchor 和 Target),并对它们的相似度进行评分,分数从 0(完全不相似)到 1(意思相同)。在数据集中,相似度(Score)表示 2 个专利短语(Anchor 和 Target)在上下文(Context)中的相似度,其值在 0~1 范围内,增量为 0.25。本文使用第 4.2 节的方法得到最终的扩展标签词,如表 2 所示,具体描述如下:

- 1)锚短语(Anchor):第 1 个专利短语。
- 2)目标短语(Target):第 2 个专利短语。
- 3)CPC 分类(Context):上下文 CPC 分类,指出进行相似度评分的专利短语的主题分类。
- 4)相似度等级(Rating):相似度等级标签,来自一个或多个手工专家评级的组合。
- 5)相似度(Score):相似度评分,来自一个或多个手工专家评级的组合。

表 2 扩展标签词示例

Table 2 Examples of extended label words

相似度	标签	标签词
0.00	not related	not related, irrelevant, totally different
0.25	low	low, holonym, domain related, field-related
0.50	medium	medium, hyponym, hypernym, structural match
0.75	high	antonym, meronym, holonym, high level, domain match, synonym
1.00	very high	very high, exact, same

2.2 问题定义

本文使用余弦距离来计算 2 个专利短语的相似度,然后计算其与人工评分之间的相关系数结果,包括 Pearson 相关系数(PCC)和 Spearman 相关系数(SRC)。

3 专利短语相似度计算方法

3.1 提示学习

提示调优通过自动编码器调整将分类任务形式化为一个掩码语言模型(MLM)问题。提示调优可用于文本分类任务,构建一个含有[MASK]的模板,然后让掩码语言模型去预测[MASK]位置的单词。因此,文本分类任务被转化为一个掩码语言建模问题。

假设 M 是一个在大规模语料库上预训练的语言模型。在文本分类任务中,输入序列 $x = (x_0, x_1, \dots, x_n)$ 被分类为类别标签 $y \in Y$ 。具体来说,提示调优用模板包装输入序列,而模板是一段自然语言文本。例如,本文将 $x =$ “The similarity between patent ‘faucet assembly’ and patent ‘tap inputs’.”分类为“not related”(相似度标记为 0)或“exact”(相似度标记为 1),可以封装成:

$$x_p = [\text{CLS}] x \text{ is } [\text{MASK}] \quad (1)$$

在专利短语相似度匹配任务中,本文使用 $x_A = \text{Word}_{\text{Anchor}}$ 表示原始专利短语(Anchor),使用 $x_T = \text{Word}_{\text{Target}}$ 表示目标专利短语(Target),则 x_A 和 x_T 可划分到一个相同相似度等级的分类标签 $y \in Y$ 中。由表 1 可以看出,在专利短语数据集中,相似度是一个离散值。因此,本文将专利短语相似度计算问题转化成专利短语分类问题来解决。使用 x_A 和 x_T 替代 x ,则式(1)可以转换成:

$$x_p = [\text{CLS}] x_A \text{ and } x_T \text{ is } [\text{MASK}] \quad (2)$$

接着,预训练的语言模型 M 给出了词汇表中每个单词 v 被 [MASK] 词元(token)的概率,即 $P_M([\text{MASK}] = v \in V_{y|x_p})$ 。在提示学习中,本文

需要一个标签词映射(verbalizer),将[MASK]位置上对于词表中词汇的预测转化成分类标签。例如在{0: “not related”, 1: “exact”}这个映射下,预训练模型在[MASK]位置,对于“not related”或“exact”这个相似程度度标签词的预测概率值会被当成是对“0”或“1”这个标签的预测值。标签词映射器首先要定义词表中哪些词是合适[MASK]词元位置的标签词,其次要定义标签词概率如何转化为最终的类别概率。

为了将专利单词的概率映射到标签的概率,本文将标签词映射定义为从词汇表中的几个单词(构成标签词集 V)到相似度标签空间 Y 的映射 g ,即 $g: v \rightarrow Y$ 。本文使用 V_y 来表示映射到特定标签 y 的 V 的子集, $\bigcup_{y \in Y} V_y = V$,则标签的概率 $P(y \in Y | x_p)$ 计算公式为:

$$P(y \in Y | x_p) = f(P_M([MASK] = v \in V_y | x_p)) \quad (3)$$

提示文本 [CLS] The similarity is [MASK] between x_p and x_t in Context [SEP] $r_A r_T$

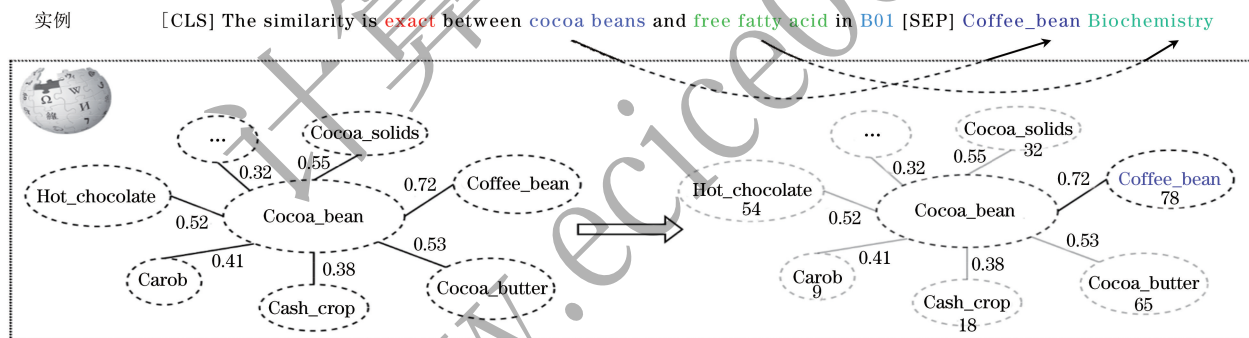


图 1 本文模型框架

Fig.1 The framework of the model in this paper

3.2 知识注入

基于上下文的掩码词预测过程并不是一对一的选择过程。因此,标签词映射中的标签词应该具备 2 个属性,即广泛的覆盖范围和较低的主观偏见。为此,本文使用外部知识来构建标签词映射,即通过 Wikidata(<https://www.wikidata.org>)获取相关实体的信息。

对于每个专利短语中的实体,本文首先使用专利短语或者标签词称作为查询关键词,通过 Wikidata 获取相关实体的信息;然后使用相似度算法对专利短语和 Wikidata 实体进行匹配,以确定它们之间的关系;最后根据匹配结果将专利短语中的实体链接到 Wikidata 知识图谱中的相应实体上,实现实体的语义统一。

为了获取专利实体间的结构信息,首先获取专利的低维嵌入表示 $\bar{x} = \text{Emb}(x)$,其中,对于

其中: f 是一个函数,其将标签词的概率值转化为标签的概率值。

在上面的例子中,常规提示调优可以定义 $V_1 = \{\text{"exact"}\}$, $V_2 = \{\text{"not related"}\}$, f 为恒等函数,如果“exact”的概率大于“not related”,则本文将这个实例划分为“exact”标签。

提示学习的本质是设计一个比较契合上游预训练任务的模板,通过增加提示模板,将训练数据转成自然语言的形式,并在合适的位置 MASK。提示学习主要包括设计预训练语言模型、设计输入模板样式和设计标签样式及模型输出映射到标签(label)的方式 3 个步骤。本文提出知识注入的提示学习方法,利用外部知识来提高提示调优的语言表达能力,其模型框架如图 1 所示,在本文方法中,通过扩展标签词及输入专利短语,达到注入外部知识的目的。

$\text{Emb}(\cdot)$, 本文方法采用的是 node2vec^[32]图嵌入算法(可更换为其他图嵌入算法)。如果在专利短语实体的嵌入之间进行相似度计算,容易忽略专利短语间的上下文信息。因此,本文对专利短语实体进行邻域过滤。为了减少邻域过滤期间的噪声,本文计算专利实体 $\bar{x}$ 与其一阶邻居实体 $N_{\bar{x}}$ 间的余弦相似度:

$$p_{x_k} = \cos(\bar{x}, x_k) = \frac{\sum_{i=1}^n \bar{x} \cdot x_k}{\sqrt{\sum_{i=1}^n \bar{x}^2} \times \sqrt{\sum_{i=1}^n x_k^2}}, x_k \in N_{\bar{x}} \quad (4)$$

如果将专利短语实体的所有一阶邻居作为知识注入的提示学习模板,则工作量巨大。然而,仅针对相似度值进行过滤,则容易丢失专利短语实体间的潜在关系。因此,本文利用邻居实体的影响度对专利短语的邻域信息过滤进行干预筛选,得到其邻域

信息,并作为知识注入模板的标签词:

$$\tilde{N}_x = \{\Phi(p_{x_k}, D_{x_k})\} \quad (5)$$

其中: D_{x_k} 为实体节点 x_k 的度; $\Phi(\cdot)$ 为邻域节点的筛选函数。本文取邻居相似度值与邻居节点度的乘积的前 m 个最大值(为了避免由于引入高阶邻居而带来的噪声影响,同时简化计算,设 $m=1$)。

如图 1 所示,以提示生成的输入文本“[CLS] The similarity is exact between cocoa beans and free fatty acid in B01 [SEP] Coffee _ bean Biochemistry”为例,目的是将专利短语对“cocoa beans”和“free fatty acid”归类到“exact”标签(即相似度为 1)。对于专利短语“cocoa beans”,首先找到它的 Wikidata 实体“Cocoa _ bean”;然后通过 Wikidata 实体间的关系构建其对应的一阶邻居图;接着采用 node2vec 图嵌入算法分别计算其与一阶邻居的相似度,如“Cocoa _ bean”与一阶邻居“Coffee _ bean”的相似度为 0.72;随后逐个计算“Cocoa _ bean”一阶邻居的度(“Coffee _ bean”一阶邻居的度为 78);最后分别计算“Cocoa _ bean”的一阶邻居相似度值与邻居节点度的乘积,将其降序排序,找出前 m 个最大的一阶邻居(设 $m=1$,即获取到最终的外部知识图谱 Wikidata 中的实体“Coffee _ bean”)。标签词“exact”与另一个专利短语“free fatty acid”也使用上述方法获取。

3.3 提示生成

本文构造有遮蔽的提示文本,其中, x_A 和 x_T 表示专利短语, r_A 和 r_T 分别为专利短语 x_A 和 x_T 的相关词, Context 为 CPC 分类, [MASK] 为预测的分类, [CLS] 和 [SEP] 为预训练模型中的分隔符。考虑到不同提示生成的文本对最终结果的影响不同,本文构造如下 $T_1 \sim T_7$ 提示生成的输入文本:

$$T_1 = [\text{CLS}] [\text{MASK}] [\text{SEP}] x_A [\text{SEP}] x_T$$

$$T_2 = [\text{CLS}] \text{ The similarity is } [\text{MASK}] \text{ between } x_A \text{ and } x_T$$

$$T_3 = [\text{CLS}] \text{ The similarity is } [\text{MASK}] \text{ between } x_A \text{ and } x_T \text{ in Context}$$

$$T_4 = [\text{CLS}] \text{ The similarity is } [\text{MASK}] \text{ between } x_A \text{ and } x_T [\text{SEP}] r_A r_T$$

$$T_5 = [\text{CLS}] \text{ The similarity is } [\text{MASK}] \text{ between } x_A \text{ and } x_T \text{ in Context } [\text{SEP}] r_A r_T$$

$$T_6 = [\text{CLS}] \text{ The similarity is } [\text{MASK}] \text{ between } x_A \text{ and } x_T [\text{SEP}] r_A r_T$$

$$T_7 = [\text{CLS}] r_A r_T [\text{SEP}] \text{ The similarity is } [\text{MASK}] \text{ between } x_A \text{ and } x_T \text{ in Context}$$

提示文本生成是基于提示学习机制构造融合专利领域知识的提示文本。添加提示文本不仅有利于模型获得上下文的语义联系,而且能够显著提升下游的任务效果,充分利用预训练语言模型可以从海量的语料中学习到通用的语言表示。

3.4 分类

本文采取对每个类所有标签词的概率值取平均的方法,定义标签词和类别的映射关系,得到最终类别 y 对应的预测概率:

$$y = \underset{y \in Y}{\operatorname{argmax}} \frac{\sum_{v \in V_y} p([\text{MASK}] = vx_p)}{|V_y|} \quad (6)$$

3.5 损失函数

在标签词映射器模块的最后阶段,本文得到一对专利短语类别的预测概率 $P(\hat{y})$ 。本文希望能够最小化真实分布与模型输出分布之间的距离,等价于最小化两者的交叉熵。交叉熵损失可以保证专利短语的分类能力,本文由此可以计算预测值与真实标签 y 之间的交叉熵损失。最终使用交叉熵损失进行分类器训练,交叉熵损失定义为:

$$= \text{CrossEntropy}(y, \hat{y}) = - \sum_{k=1}^K y_k \log_a \hat{y}_k \quad (7)$$

其中: y 表示数据集中的真实相似度标签; $\hat{y}$ 表示本文模型预测的相似度类别标签。

4 实验验证

4.1 实验设置

在实验过程中,本文模型和对比基线都是通过 Python 来实现,使用工具包 OpenPrompt(<https://github.com/thunlp/OpenPrompt>)实现基于知识注入的专利短语相似度分类方法。Our(T_5)是本文提出的基于知识注入提示学习的专利短语相似度计算方法。使用性能最好的提示文本 T_5 与基线方法进行比较。

4.2 对比方法

为了证明本文方法的有效性,将所提方法与经典方法及最近出现的方法进行比较,对比方法具体如下:

1) Word2Vec^[33] 从大量文本预料中以无监督方式学习语义知识,用于学习文本中的单词表示。本文使用 tensorflow_hub 工具包加载英语维基百科语料版的 Word2Vec 预训练文本嵌入向量。

2) GloVe^[34] 构造一个单词共现矩阵,然后最小化词向量之间的欧几里得距离的平方来学习词向

量。本文通过加载预训练好的“Wikipedia2014 + Gigaword(glove, 6B, zip)”词向量实现 GloVe。

3) FastText^[35] 基于用实数向量表示文本中单词的思想,使用神经网络学习从单词到向量的映射。本文加载预训练好的“wiki-news-300d-1M. vec”词向量获取每个专利短语的 FastText 向量表示。

4) BERT^[7] 和 PatentBERT。BERT 是一个基于 Transformer 的模型,使用一个大型的无监督语言建模数据集来学习一个句子的表示,这个句子中单词的顺序是不变的。对于 BERT,本文使用 BERT-large 模型。为了进行比较,本文在与 BERT-large 相同大小的专利数据上进行训练后得到 PatentBERT 模型。

5) Sentence-BERT^[36] 使用深度学习模型来预测句子中的下一个单词。本文使用“sentence-transformers”工具包加载“all-mpnet-base-v2”实现 Sentence-BERT。

6) P-tuning^[31] 利用可训练的向量来自动构建提示文本模板,使用 BiLSTM 对模板中的初始化向量进行表征,增加嵌入向量之间的相关性。

7) KPT^[12] 通过外部知识库扩展描述器的提示调优,用于提高文本分类的准确率。

8) Patent2Vec^[18] 用于专利分类,使用图嵌入来生成低维表示,通过视图增强模块和基于注意力的多视图融合方法来丰富和对齐不同视图的信息。

4.3 评估指标

在语义相似性计算任务中,给定一个数据集,可以通过计算预测得分与人工评分之间的相关系数来评价模型性能。有 2 种常用的相关系数,即 PCC 和 SRC,前者用来衡量 2 个变量之间的关系强度,后者偏向于相关度的评价。

除此之外,本文还使用文本分类评估指标进行比较,其中包括准确率($R_{Accuracy}$)、精确率($R_{Precision}$)、召回率(R_{Recall})和 F1 值,它们的结果包括 4 个部分,即真正例(N_{TP})、假正例(N_{TN})、真负例(N_{FP})和假负例(N_{FN}),则对应的 $R_{Accuracy}$ 、 $R_{Precision}$ 、 R_{Recall} 和 F1 值的计算公式分别为:

$$R_{Accuracy} = \frac{N_{TP} + N_{TN}}{N_{TP} + N_{FP} + N_{FN} + N_{TN}} \times 100\%$$

$$R_{Precision} = \frac{N_{TP}}{N_{TP} + N_{FP}} \times 100\%$$

$$R_{Recall} = \frac{N_{TP}}{N_{TP} + N_{FN}} \times 100\%$$

$$F_1 = 2 \times \frac{R_{Precision} \times R_{Recall}}{R_{Precision} + R_{Recall}} \times 100\%$$

4.4 实验结果

表 3 所示为对比方法在专利短语相似度匹配数据集 PPSD 上的 PCC 和 SRC 性能表现。从表 3 可以看出:静态预训练模型 Word2Vec、GloVe 和 FastText 性能表现不佳,考虑到数据集结构(如许多具有不同含义的匹配专利术语)的影响,实验结果符合预期;PatentBERT 模型显著优于常规的 BERT 模型,这意味着通用预训练模型对于专利中发现的专利术语而言效果不佳。

表 3 不同语义相似度计算方法在 PPSD 上的性能表现

Table 3 Performance of different semantic similarity calculation methods on PPSD

方法	PCC	SRC
Word2Vec	0.439	0.488
GloVe	0.431	0.449
FastText	0.412	0.467
BERT	0.419	0.411
PatentBERT	0.532	0.538
Sentence-BERT	0.597	0.579
P-tuning	0.648	0.631
KPT	0.672	0.668
Patent2Vec	0.668	0.656
Our(T_5)	0.718	0.706

然而,本文从 PatentBERT 和 Sentence-BERT 模型中得到了一个较好的结果,因为已经针对用于短语相似性匹配数据的预训练进行了微调。KPT 和本文方法都是基于提示学习进行微调的提示调优方法,表 3 显示,本文专利短语相似度计算方法(Our(T_5))性能优于 KPT,也比所有的对比基线方法更优,这验证了将知识图谱中蕴含的知识引入提示学习中的数据增强方式具有有效性。

4.5 实验分析

4.5.1 提示文本的影响

表 4 所示为使用不同提示文本的 PCC、SRC、准确率、精确率、召回率、F1 值的实验结果,最优结果加粗标注。表 4 显示,提示文本 $T_1 \sim T_7$ 的 PCC 波动比较明显,在 T_5 上取得了最好的 PCC 性能。从式(6)可知, T_5 考虑了上下文 Context 特征信息和相关词 relatedWord 特征信息,说明本文计算方法具有有效性,同样的结论也可从 T_5 在 SRC 上的性能表现得出。

从表 4 可以看出: $T_1 \sim T_7$ 的准确率波动较小,在 T_4 和 T_5 上取得了较好的准确率,可知本文方法对正样本的查准效果更好; $T_1 \sim T_7$ 的精确率波动也较小,在 T_5 上取得了最佳值,反映了本文方法分

表 4 不同提示文本的实验结果

Table 4 Experimental results of different prompt texts

提示文本	PCC	SRC	$R_{Accuracy}$	$R_{Precision}$	R_{Recall}	F_1
T_1	0.674	0.661	0.669	0.669	0.647	0.658
T_2	0.678	0.662	0.664	0.654	0.644	0.649
T_3	0.685	0.676	0.673	0.663	0.633	0.648
T_4	0.681	0.680	0.685	0.674	0.647	0.660
T_5	0.718	0.706	0.683	0.680	0.643	0.661
T_6	0.705	0.696	0.668	0.658	0.648	0.653
T_7	0.689	0.688	0.680	0.679	0.640	0.658

类结果准确率稳定,也说明整体分类效果较好;召回率是正确预测专利短语分类数与数据集 PPSD 中该类别实际包括专利数的比值,在 $T_1 \sim T_7$ 上表现相当稳定,说明本文方法对正样本的查全效果较好;F1 值是精确率和召回率的调和平均数,综合平衡了精确率和召回率,在 $T_1 \sim T_7$ 上的表现波动更小,说明本文计算方法的整体分类效果较好。

图 2 所示为不同提示文本在验证集上的损失值变化情况(彩色效果见《计算机工程》官网 HTML 版)。从图 2 可以看出,在 Epoch 为 5 之前,不同提示文本的验证集损失值波动较大,但是随后就稳定下来,说明模型泛化效果较好,也进一步说明了其有效性。

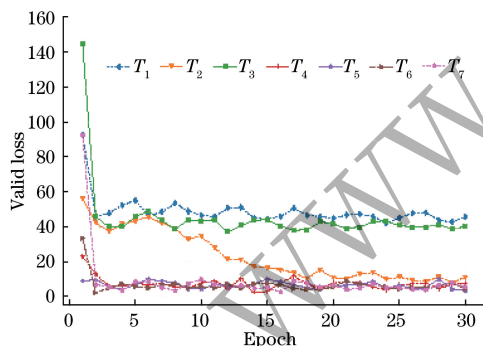


图 2 不同提示文本在验证集上的损失值变化情况

Fig.2 Changes in loss values of different prompt texts on the validation set

4.5.2 消融实验分析

从式(6)的提示文本 $T_1 \sim T_7$ 中选出提示文本 T_1 、 T_2 、 T_4 和 T_5 进行消融实验分析。从表 5 的实验结果可以看出,在提示阶段,不同的提示文本对应的 PCC 和 SRC 不同且相差较大,提示文本的长度、上下文 Context 和相关词的选择均会影响最终的实验结果,且提示文本中不同特征信息交换位置时也在一定程度上影响了拼接后的语义,导致 PCC 和 SRC 较低。因此,寻求合适的提示文本是提示学习能否取得较好性能的关键。

表 5 考虑不同特征信息时提示文本的

PCC 和 SRC 性能表现

Table 5 PCC and SRC performance of prompt text considering different feature information

提示文本	考虑的特征信息			PCC	SRC
	Anchor, Target, Score	Context	relatedWord		
T_1	✓			0.674	0.661
T_2	✓	✓		0.678	0.662
T_4	✓		✓	0.681	0.680
T_5	✓	✓	✓	0.718	0.706

对于提示文本 T_1 、 T_2 、 T_4 和 T_5 ,在提示文本中输入的合理信息越多,其性能越优,该结果也能验证本文基于知识注入提示学习的专利短语相似度计算方法的有效性。

对比提示文本 T_1 和 T_2 可以看出, T_2 的 PCC 和 SRC 性能表现都优于 T_1 ,说明 T_2 的提示文本设置比 T_1 更加合理,也说明了上下文 Context(CPC 分类)的有效性。对比 T_4 和 T_5 可知,缺失 CPC 分类特征信息对 T_4 的 PCC 和 SRC 性能影响更大,也进一步验证了上下文 Context 的重要性。

对比提示文本 T_1 和 T_4 , T_4 的 PCC 和 SRC 表现优于 T_1 ,说明 T_4 的提示文本设置比 T_1 更加合理,也说明了相关词 relatedWord 特征信息的有效性。对比 T_2 和 T_4 可知, T_4 的 PCC 和 SRC 性能更佳,说明考虑相关词 relatedWord 特征信息比考虑上下文信息更加有效,也验证了相关词筛选方法的有效性。相比 T_2 , T_5 的 PCC 和 SRC 性能大幅提升,说明了相关词 relatedWord 特征信息的有效性,也验证了相关词选择方法的合理性。

从表 5 还可以看出,不同提示文本之间的 PCC 和 SRC 性能表现差距较大。相比 T_1 , T_5 的 PCC 和 SRC 性能大幅提升,说明上下文 Context 信息和相关词 relatedWord 特征信息具有有效性,进一步验证了本文计算方法的合理性。

5 结束语

专利是迈向知识型社会的重要资源,开发高效的系统来管理海量专利数据非常重要。为了解决专利短语信息不足的问题,本文利用专利短语中的相似度标签信息,提出一种基于知识注入的提示学习方法。实验结果验证了该方法的有效性。随着人工智能神经网络技术的不断发展,基于数据驱动的预训练和提示学习技术变得更加有效,但是人类的知识也在不断更新,如在不同时间阶段三元组知识(苹果,

CEO,乔布斯)会更新为(苹果,CEO,库克)。因此,更新预训练语言模型中的参数空间知识将是下一步的研究方向。

参考文献

- [1] 刘小蝶,朱筠,晋耀红. 中文专利中有标记并列结构的自动识别研究[J]. 计算机工程, 2018, 44(6): 162-168, 175.
LIU X D, ZHU Y, JIN Y H. Research on automatic identification of marked parallel structures in Chinese patent [J]. Computer Engineering, 2018, 44(6): 162-168, 175. (in Chinese)
- [2] SOUSA M J, JAMIL G, WALTER C E, et al. Big data analytics on patents for innovation public policies[J]. Expert Systems, 2023, 40(1): e12673.
- [3] 罗艺雄,吕学强,游新冬. 融合多特征的专利功效短语识别[J]. 中文信息学报, 2022, 36(12): 139-148.
LUO Y X, LÜ X Q, YOU X D. Patent efficacy phrase recognition based on multiple features[J]. Journal of Chinese Information Processing, 2022, 36(12): 139-148. (in Chinese)
- [4] 殷亚珏,高晓雅,王晶晶,等. 基于多视角注意力机制的专利匹配方法[J]. 中文信息学报, 2022, 36(7): 106-113.
YIN Y J, GAO X Y, WANG J J, et al. Patent matching with multi-view attentive network[J]. Journal of Chinese Information Processing, 2022, 36(7): 106-113. (in Chinese)
- [5] KIM J, YOON J, PARK E, et al. Patent document clustering with deep embeddings[J]. Scientometrics, 2020, 123(2): 563-577.
- [6] MASE H, MATSUBAYASHI T, OGAWA Y, et al. Proposal of two-stage patent retrieval method considering the claim structure[J]. ACM Transactions on Asian Language Information Processing, 2005, 4(2): 190-206.
- [7] DEVLIN J, CHANG M W, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of Conference of the North American Chapter of the Association for Computational Linguistics. [S. l.]: Association for Computational Linguistics, 2019:4171-4186.
- [8] BINZ M, SCHULZ E. Using cognitive psychology to understand GPT-3[J]. Proceedings of the National Academy of Sciences of the United States of America, 2023, 120(6): e2218.
- [9] ALT C, GABRYSAK A, HENNIG L. TACRED revisited: a thorough evaluation of the TACRED relation extraction task[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. [S. l.]: Association for Computational Linguistics, 2020: 1558-1569.
- [10] 张宁豫,谢辛,陈想,等. 基于知识协同微调的低资源知识图谱补全方法[J]. 软件学报, 2022, 33(10): 3531-3545.
ZHANG N Y, XIE X, CHEN X, et al. Knowledge collaborative fine-tuning for low-resource knowledge graph completion[J]. Journal of Software, 2022, 33(10): 3531-3545. (in Chinese)
- [11] BHARGAVA P, NG V. Commonsense knowledge reasoning and generation with pre-trained language models: a survey [J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2022, 36(11): 12317-12325.
- [12] SINGH G V, FIRDAUS M, MISHRA S, et al. Knowing what to say: towards knowledge grounded code-mixed response generation for open-domain conversations [J]. Knowledge-Based Systems, 2022, 249: 108900.
- [13] HU S D, DING N, WANG H D, et al. Knowledgeable prompt-tuning: incorporating knowledge into prompt verbalizer for text classification[C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. [S. l.]: Association for Computational Linguistics, 2022: 2225-2240.
- [14] YE H B, ZHANG N Y, DENG S M, et al. Ontology-enhanced prompt-tuning for few-shot learning [C] // Proceedings of the ACM Web Conference. New York, USA: ACM Press, 2022: 778-787.
- [15] FUJII A. Enhancing patent retrieval by citation analysis[C]// Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, USA: ACM Press, 2007: 793-794.
- [16] KANG I S, NA S H, KIM J, et al. Cluster-based patent retrieval[J]. Information Processing & Management, 2007, 43(5): 1173-1182.
- [17] LI S B, HU J, CUI Y X, et al. DeepPatent: patent classification with convolutional neural networks and word embedding[J]. Scientometrics, 2018, 117(2): 721-744.
- [18] FANG L T, ZHANG L, WU H, et al. Patent2Vec: multi-view representation learning on patent-graphs for patent classification[J]. World Wide Web, 2021, 24(5): 1791-1812.
- [19] GUU K, LEE K, TUNG Z, et al. Retrieval augmented language model pre-training[C]//Proceedings of International Conference on Machine Learning. New York, USA: ACM Press, 2020: 3929-3938.
- [20] LIU B, LIN T, LI M. Enhancing aspect-category sentiment analysis via syntactic data augmentation and knowledge enhancement[J]. Knowledge-Based Systems, 2023, 264: 110339.
- [21] INCITTI F, URLI F, SNIDARO L. Beyond word embeddings: a survey[J]. Information Fusion, 2023, 89: 418-436.
- [22] 陈彦桦,李剑. 一种基于结构特征的树相似度计算方法[J]. 计算机工程, 2018, 44(11): 197-201, 208.
CHEN Y H, LI J. A tree similarity computation method based on structure feature[J]. Computer Engineering, 2018, 44(11): 197-201, 208. (in Chinese)
- [23] ROGERS A, GARDNER M, AUGENSTEIN I. QA dataset explosion: a taxonomy of NLP resources for question answering and reading comprehension[J]. ACM Computing Surveys, 2023, 55(10): 1-45.
- [24] 陆晓蕾,倪斌. 基于预训练语言模型的BERT-CNN多层级专利分类研究[J]. 中文信息学报, 2021, 35(11): 70-79.
LU X L, NI B. BERT-CNN: a hierarchical patent classifier based on pre-trained language model[J]. Journal of Chinese Information Processing, 2021, 35(11): 70-79. (in Chinese)
- [25] 亢晓勉,宗成庆. 基于篇章结构多任务学习的神经机器翻译[J]. 软件学报, 2022, 33(10): 3806-3818.
KANG X M, ZONG C Q. Neural machine translation based on multi-task learning of discourse structure[J]. Journal of Software, 2022, 33(10): 3806-3818. (in Chinese)
- [26] LUNDB D, WANG T. Chatting about ChatGPT: how may AI and GPT impact academia and libraries? [J]. Library Hi Tech News, 2023, 40(3): 26-29.
- [27] SONG C Y, CAI F, WANG M R, et al. TaxonPrompt: Taxonomy-aware curriculum prompt learning for few-shot event classification [J]. Knowledge-Based Systems, 2023, 264: 110290.
- [28] LIU X, JI K X, FU Y C, et al. P-tuning: prompt tuning can be comparable to fine-tuning across scales and tasks[C]// Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. [S. l.]: Association for Computational Linguistics, 2022: 61-68.
- [29] SCHICK T, SCHÜTZE H. Exploiting cloze-questions for few-shot text classification and natural language inference [C]//Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics. [S. l.]: Association for Computational Linguistics, 2021:

- 255-269.
- [30] LI X L, LIANG P. Prefix-tuning: optimizing continuous prompts for generation[C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. [S. l.]: Association for Computational Linguistics, 2021: 4582-4597.
- [31] SUN T X, LIU X Y, QIU X P, et al. Paradigm shift in natural language processing [J]. Machine Intelligence Research, 2022, 19(3): 169-183.
- [32] GROVER A, LESKOVEC J. node2vec: scalable feature learning for networks[C]//Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA: ACM Press, 2016: 855-864.
- [33] MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distributed representations of words and phrases and their compositionality [EB/OL]. [2023-04-05]. <https://arxiv.org/abs/1310.4546>.
- [34] PENNINGTON J, SOCHER R, MANNING C. GloVe: global vectors for word representation[C]// Proceedings of 2014 Conference on Empirical Methods in Natural Language Processing. [S. l.]: Association for Computational Linguistics, 2014: 1532-1543.
- [35] JOULIN A, GRAVE E, BOJANOWSKI P, et al. Bag of tricks for efficient text classification[C]// Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics. [S. l.]: Association for Computational Linguistics, 2017: 427-431.
- [36] REIMERS N, GUREVYCH I. Sentence-BERT: sentence embeddings using siamese BERT-networks[C]//Proceedings of 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. [S. l.]: Association for Computational Linguistics, 2019: 3982-3992.

编辑 吴云芳