小样本数据分类任务

痛！太痛了！

李一鸣  
 计算机科学与技术学院&2022级  
 中国科学院大学  
 中国-北京  
liyiming22s1@ict.ac.cn.

张兆  
 计算机科学与技术学院&2022级  
 中国科学院大学  
 中国-北京 zhaozhao809@163.com

李想  
 人工智能学院&2022级  
 中国科学院大学  
 中国-北京  
2300049883@qq.com

赵家乐  
 计算机科学与技术学院&2022级  
 中国科学院大学  
 中国-北京  
 1007613549@qq.com

团队简介

在此处插入段落文字。 在此处插入段落文字。 在此处插入段落文字。 在此处插入段落文字。 在此处插入段落文字。 在此处插入段落文字。 在此处插入段落文字。 在此处插入段落文字。

在此处插入段落文本。 在此处插入段落文字。 在此处插入段落文字。 在此处插入段落文字。 在此处插入段落文字。 在此处插入段落文字。 在此处插入段落文字。 在此处插入段落文字。 在此处插入段落文字。 在此处插入段落文字。 在此处插入段落文字。

在此处插入段落文字。 在此处插入段落文字。 在此处插入段落文字。 在此处插入段落文字。 在此处插入段落文字。 在此处插入段落文字。

在此处插入段落文字。 在此处插入段落文字。 在此处插入段落文字。 在此处插入段落文字。 在此处插入段落文字。 在此处插入段落文字。

摘要

近年来，随着政策扶植、国家工业化水平和国民教育水平提高，我国的专利申请量爆发式增长。因此对于专利文本的分类管理需求持续增加。得益于自然语言处理技术的发展，大规模预训练模型在学术界与工业界被广泛应用。在小样本的专利文本分类问题中，我们充分利用ERNIE 3.0模型，设计了一个基于伪标签生成与数据增强的小样本文本分类方法，能够很好地适应小样本学习场景。实验结果表明，在CCF BDCI的小样本数据分类任务上，我们的方法最终在B榜取得了0.5955的F1分数。

关键词

专利文本，文本分类，小样本，伪标签。

引言

近年来，随着我国科技的进步和创新能力的提升，我国的专利申请量快速增长，专利检索、查新、管理等需求也不断增加。为了满足以上需求，提升专利服务质量，通常需要建立多个维度的专利分类体系。常见的分类体系有国际专利分类(IPC)、联合专利分类(CPC)、欧洲专利分类(ECLA)等，但是这些分类体系比较复杂，专业性强，对非IP人员而言使用有一定的困难。近年来，随着大规模与训练模型的发展，使得自然语言处理技术在工业界的应用成为可能。也为解决专利分类问题带来了而新机遇。

相较于传统的文本，专利文本的数据量较为稀少，因此，小样本学习在专利文本分类中至关重要。此外，不同领域的专利数量大不相同，某些类别的样本数量明显少于其他类别，存在长尾分布的问题，传统的小样本学习方法并不能很好地预测长尾类别。

为了解决上述问题，我们提出了基于伪标签生成与数据增强的小样本文本分类方法。在我们的方案中，同时考虑到小样本学习问题以及数据集中存在长尾分布的问题。与传统伪标签处理方法，预选选定阈值不同，我们为每一个类别 设置一个单独的阈值 。进一步地，为了缓解数据不平衡的问题，我们为数据集中的尾部类别应用2种数据增强方法。实验结果表明，我们的方法在小样本专利文本分类的任务中，取得了很好的效果。在初赛A榜测试集上取得了高达0.6516的F1分数，在B榜测试集上取得了高达0.5955的F1分数。

1 任务描述

本赛题要求对小样本的专利文本数据进行分类，包括958条专利数据，包括专利权人、专利标题、专利摘要和分类标签，其中标签经过脱敏处理，共36类。

本赛题采用Macro-F1值，具体定义如下：

其中，TP为真正例，FP为假正例，FN为假负例，通过上述公式得到每一类的F1值，再对每一类的F1值求平均，即得到最终的Macro-F1值。

2 解决方案

针对小样本专利文本分类问题，我们提出了一种伪标签生成方法来扩充训练集，解决小样本分类中的数据稀缺难题。图1展示了我们解决方案的全部流程图。具体地，首先我们根据模型输出的softmax分数为不同类别的样本设置不同阈值，来生成伪标签扩充训练集规模。接着，使用扩充后的训练集对ernie-3.0-base语言模型在文本分类任务上进行微调。此外，为了提升模型的鲁棒性以及解决类别不平衡问题问题，我们进一步采用了对抗训练，提升模型的鲁棒性，实现了更加精准的分类；并针对尾部样本进行数据增强，缓解了尾部类别的数据稀疏问题。

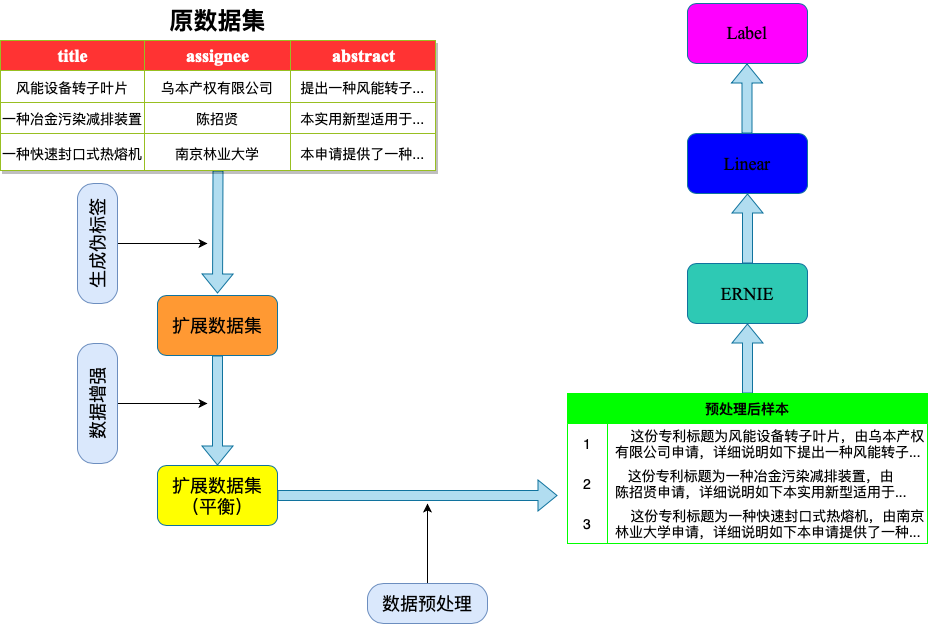


图1：解决方案的整体流程图

2.1 数据预处理

原数据集的每个样本包含title、assignee、abstract字段，这些信息包含不同粒度的信息，为了更好地将各字段组织起来，用于文本分类，我们受近年在自然语言处理领域中新兴的提示学习启发，将各字段重新组织为一段文本，基于以下模板构造模型的输入。

“这份专利的标题为《{x.title}》，由{x.assignee}公司申请，详细说明如下：{x.abstract}”

2.2 对抗训练

对抗训练旨在提升模型的鲁棒性，同时也能提高模型的泛化能力。通常是对原始输入样本 添加扰动 ，得到对抗样本 ，再将对抗样本输入模型训练。

具体地，我们采用FGM (Fast Gradient Method)攻击算法，其所施加的扰动与样本的梯度方向相同：

新的对抗样本为

2.3 伪标签

这里介绍伪标签产生流程

2.4 数据增强

为了缓解类别不平衡的问题，我们对少数类（长尾类别12、22、32、35）应用两种数据增强方法：

回译：我们利用谷歌翻译接口，将中文翻译为英法德日韩5种语言，再翻译回中文，得到新的样本。

EDA：Easy Data Augmentation，由4种不同的方法组成，分别是同义词替换（从句子中随机选择n个非停用词，随机选择它们的同义词进行替换）、随机插入（从句子中随机选择一个非停用词的单词，随机选择它的一个近义词并将它插入在句子的任意位置。并将此过程重复n次）、随机替换（随机选择句子中的两个单次并将它们交换位置。将此过程重复n次）与随机删除（对句子中的每一个单词，都以一给定概率p判定此单词是否被删除）。

3 实验结果

在此处插入段落文本。 在此处插入段落文字。 在此处插入段落文字。 在此处插入段落文字。 在此处插入段落文字。 在此处插入段落文字。 在此处插入段落文字。 在此处插入段落文字。 在此处插入段落文字。 在此处插入段落文字。 在此处插入段落文字。在此处插入段落文字。 在此处插入段落文字。 在此处插入段落文字。

致谢

在此处插入段落文本。 在此处插入段落文字。 在此处插入段落文字。 在此处插入段落文字。 在此处插入段落文字。 在此处插入段落文字。 在此处插入段落文字。 在此处插入段落文字。 在此处插入段落文字。 在此处插入段落文字。 在此处插入段落文字。在此处插入段落文字。 在此处插入段落文字。 在此处插入段落文字。

参考

[1] 作者（2018年）。 文章标题。 期刊名称，10（3），1-10。

[2] 作者，文章标题，https：//www.bdci.cn/

[3] 作者1和作者2，2018年文章标题，第二册的标题（第2版）。 出版商名称，城市，省份，国家。

[4] 编辑姓名（编辑）。 2018.第二册的标题（第二版）。 XXX大学出版社，城市，第100章. https：//www.bdci.cn/

**注意：**

1. 所有团队论文（方案说明）可能会统一印刷程作品集，并做宣传、展示使用。
2. 本文需严格按照样例格式（字体、字号、间距等）排版，若有未说明清楚部分，请参照期刊论文格式，保证美观大方，不得少于3页（第3页至少大于半页），不得大于6页；
3. 请仔细检查，文本准确、得体，不要出现错词、错句，注意删除和本论文无关的官方备注、说明用内容。

会议名称：ACM伍德斯托克会议

会议简称：WOODSTOCK'18

会议地点：美国德克萨斯州埃尔帕索

ISBN：978-1-4503-0000-0 / 18/06

年份：2018

时间：6月

版权年份：2018年

版权声明：维护权利

DOI：10.1145 / 1234567890

RRH：F。Surname等。

价格：$ 15.00