基于特征选择和文心模型的文本分类实践

——中文期刊论文自动分类技术报告

参赛队伍：海思路100号

成员：孟晓龙 任正非

CONTENTS

01. 背景与意义

现实背景

囿于人工有限的精力和图书编码的繁杂，如何利用信息技术实现中文期刊论文自动分类是图书情报领域值得研究和实践的课题。

科研意义

中文期刊论文自动分类是文本分类的子领域任务，但受限于相关优质开放数据较少、数据呈现长尾分布特性等原因，使之成为研究相对较少且极具挑战的任务。

02.思路与方案-评标and评测

评标

1.国产化——百度PaddlePaddle深度学习开发框架、百度AI Studio实践平台

2.全面性——题名、关键词、刊名与摘要，和分类号

3.可推广——ERNIE 3.0 文心模型

评测

准确率——数据增强、模型选择、模型集成等

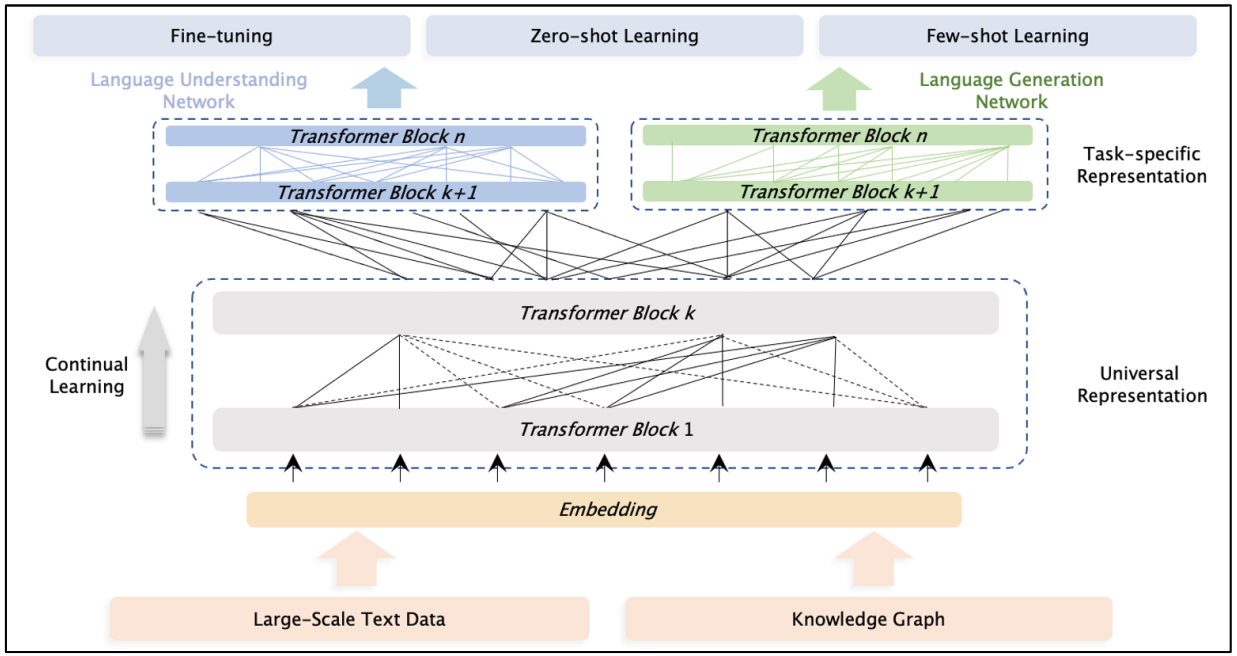
03. 要点与结果-模型基本架构和处理流程

中文期刊论文自动（四级类目）分类任务本质是具有长尾分布的单标签文本分类任务，故算法模型基本架构和处理流程如下：

1. 加载文本数据；2. 数据进行预处理，将数据转换成id的形式，放入到dataloader中，用于训练过程中给模型提供批量的数据； 3. 经过ERNIE/BERT/RoBERTa 等模型训练，得到每条文本对应每个标签（单标签多分类）的概率，通常概率最大的标签即作为该文本的标签；4. 解析成文本对应的标签。

03. 要点与结果-基线模型

本测评任务训练使用的基线模型为百度融合大规模知识的ERNIE 3.0（即文心）预训练模型，模型结构如图所示。



模型	lr_scheduler	topic+keyword+journal+abstract+	topic+keyword+journal+abstract++
bert-base-chinese	LinearDecay	0.76455	0.76186
ernie-3.0-xbase-zh	LinearDecay	0.77484	0.77356

注：因对“总分复用符号（-）”的不解，分别作为四级类目分类号（即9230/8745类别，并在后续实验记“+”和“++”以区别）。

03. 要点与结果-特征选择

给定中文期刊论文自动（四级类目）分类的测试数据，包含id、题名、关键词、刊名、摘要和分类号等数据条目，从如何合理地选择给定数据的特征来源将十分重要。分别以ERNIE 3.0 Medium 预训练模型为基础，实验“题名+摘要（记topic+ abstract）”、“题名+关键词（记topic+ keyword）”、“题名+关键词+摘要（记topic+ keyword+ abstract）”、“题名+关键词+刊名+摘要（记topic+ keyword+journal+ abstract）”，来测验特征数据来源的选取，如下表2 所示（说明：此处由于样本类别长尾分布，故label 5/10 分别表示仅选取给定数据中类别样本大于等于5/10 的数据，方便训练集和验证集4:1 划分）。

模型	label	topic+keyword+	topic+keyword++	topic+abstract+	topic+abstract++
ernie-3.0-medium-zh	5	0.69077	0.69464	0.63062	0.63293
ernie-3.0-medium-zh	10	0.69839	0.69908	0.63477	0.63675
模型	label	topic+keyword+abstract+	topic+keyword+abstract++	topic+keyword+journal+abstract+	topic+keyword+journal+abstract++
ernie-3.0-medium-zh	5	0.71106	0.71355	0.72296	0.72521
ernie-3.0-medium-zh	10	0.71722	0.71726	0.72891	0.72969

03. 要点与结果-长尾分布与数据增强

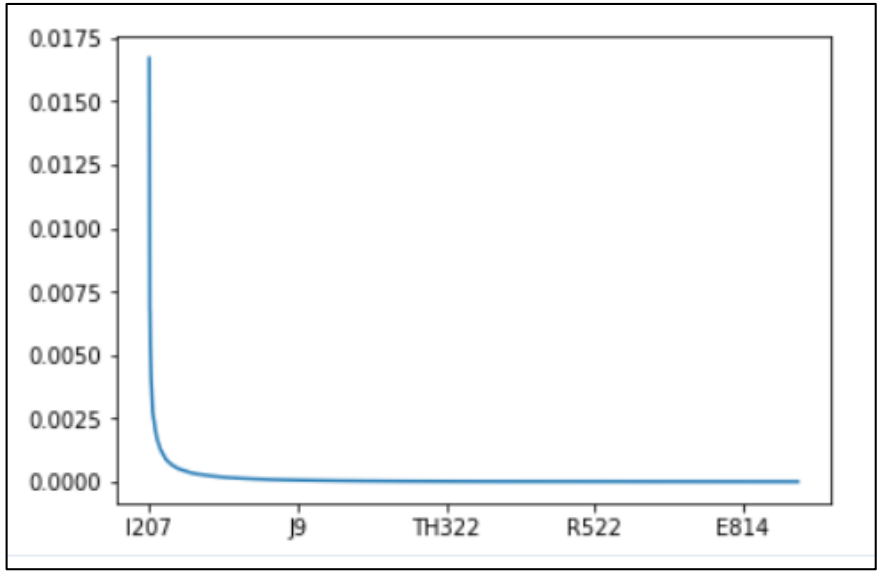


图 样本类别长尾分布情况

针对样本长尾分布的问题，参考相关文档资料，使用随机同义词替换、等价字替换、随机置换邻近的字和随机字删除等数据增强方法组合，实现长尾部分数据增加约12 倍（可调整增强文本个数create_num 和文本改变率change_rate 参数）。

模型	label	topic+keyword+journal+abstract+	topic+keyword+journal+abstract++
ernie-3.0-medium-zh	5	0.72296	0.72521
ernie-3.0-medium-zh	10	0.72891	0.72969
ernie-3.0-medium-zh	plus	0.75868	0.75752
ernie-3.0-base-zh	plus	0.76363	0.76351
ernie-3.0-xbase-zh	plus	0.77484	0.77356

03. 要点与结果-模型集成

参考相关文档资料，分别使用多结果等权投票融合（即当预测结果为具体的类别而非概率时，基于少数服从多数的原则选择集体同意的类别）和多结果平均融合（即输出结果为概率分数时，通过采用多个个体模型预测值的平均降低过拟合），其结果如下：

模型	topic+keyword+journal+abstract+	topic+keyword+journal+abstract++
多结果等权投票融合	0.78624	0.78436
多结果平均融合	待完成（不佳）	

04.启示与起点

1.数据增强，尝试如对抗训练，FGSM、PGD等；对比学习，R-Drop损失函数等。

2.损失函数修正（代价敏感学习），如正例的权重； Focal loss损失函数等。

3. “刊名”是否选择。

4. “摘要”的文本摘要生成任务。

5.多标签文本分类任务。

谢谢聆听

敬请各位老师批评指正

Speaker name and title