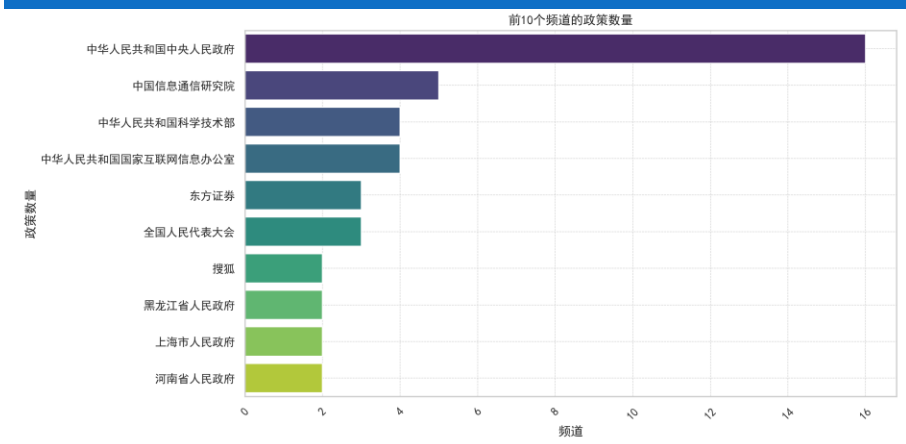
PolicyBERT: 基于全词掩码的中文政策文本预训练语言模型

北京大学 信息管理系 大数据管理与应用专业 浦皓天

研究背景

1.中文文本因缺乏空格标记和分词歧义问题，导致现有词嵌入模型难以准确建模真实语义。
2.融合词级语义信息是一个具有潜力的途径。
3.本文提出PolicyBERT模型，结合HanLP分词与注意力机制，提升中文政策文本的语义理解能力。
4.基于检索增强生成（RAG）技术，结合PolicyBERT和大语言模型，构建高效的政策文本问答系统。

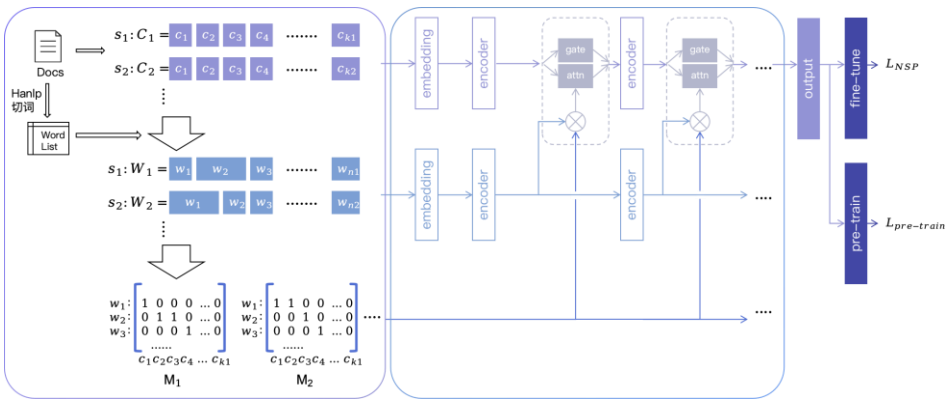
数据集-PolicySM



1.来自61个来源共99篇中央、地方政策文件。总字数超过900,000。
2.基于不同规则与任务，构建了正负样本1:1,1:5的数据集PolicySM-1和PolicySM-2，样本数分别为46748与140244

数据集	正样本数	负样本数	正负比例	总样本数
PolicySM-1	23374	23374	1:1	46748
PolicySM-2	23374	116870	1:5	140244

PolicyBERT



1.利用语料库构建词表。
2.根据词表对每个句子进行分词，计算匹配矩阵M。
3.使用单独的词编码器，与字符编码器并行，每一层输出与M相乘，使用融合层（门控或多头注意力）融合。
4.字符编码器的最终输出进入预训练头或下游任务头。

公开数据集实验结果

Model\Task	CWS	POS	NER	DC	SA	SPM
BERT	0.9720	0.9543	0.9312	0.9671	0.9410	0.8513
BERT-wwm	-	-	**0.9510**	**0.9760**	0.9500	0.8680
ERNIE 1.0	-	-	0.9510	0.9730	0.9540	0.8740
ERNIE 2.0(B)	-	-	-	-	0.9550	0.8790
NEZHA(B)	-	-	-	-	0.9517	0.8741
NEZHA-wwm(B)	-	-	-	-	0.9584	0.8710
ERNIE 2.0(L)	-	-	-	-	0.9580	0.8790
NEZHA(L)	-	-	-	-	0.9583	0.8720
NEZHA-wwm(L)	-	-	-	-	**0.9600**	0.8794
ZEN	0.9789	0.9582	0.9324	0.9687	0.9442	0.8527
$PolicyBERT_{attn}$	0.9812	0.9711	0.9437	0.9651	0.9441	**0.9016**
$PolicyBERT_{gate}$	**0.9816**	**0.9716**	0.9422	0.9655	0.9475	0.8997

PolicySM-1实验结果

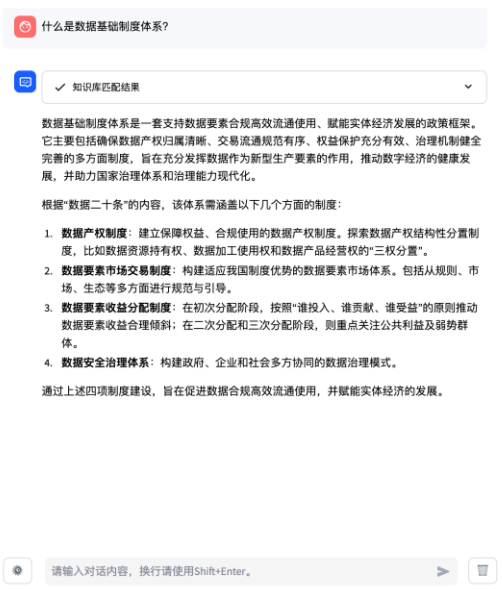
融合方法	分词方法	准确率
None	n-gram	0.8867
Gated Fusion	n-gram	0.8959
Attention-based Fusion	n-gram	**0.9121**
None	Hanlp	0.9035
Gated Fusion	Hanlp	0.8981
Attention-based Fusion	Hanlp	**0.9132**

PolicySM-2实验结果

骨干模型	融合方法	NSP acc
BGE	none	0.9236
	gate	**0.9311**
	attn	0.9272
ZEN	none	0.9260
	gate	**0.9325**
	attn	0.9274

1.在多个公开数据集上取得最优、次优结果（分词、词性标注、句对匹配）
2.在PolicySM-1，PolicySM-2数据集上验证了引入分词、融合层的有效性。

未名问政-PolicyBert



1.中文使用PolicyBERT框架对bge-base-zh-v1.5的部分权重进行微调，得到bge-merged作为文档匹配模型。
2.利用Ollama框架本地部署bge-merged和通用大语言模型千问qwen2.5:7B。
3.基于langchain-chatchat框架，部署本地政策文件知识库，搭建政策问答RAG系统。

*代码与数据集公开在github.com/PuHT4213/PolicyBert和github.com/PuHT4213/WeiMingPolicyRAG上