基于大语言模型的BCC语料库自然语言检索

刘廷超[1] 鲁鹿鸣[1] 荀恩东[2] 靳泽莹[1] 杨兆勇[1]

1. 北京语言大学 信息科学学院
2. 北京语言大学 高精尖语言资源中心

摘要

- 语料库在语言学和自然语言处理领域至关重要。北京语言大学的BCC语料库，资源丰富且检索高效，备受推崇，然而，其BCC检索式的复杂性限制了普及。为此，本文提出TextToBCC模型，目标是实现自然语言对BCC语料库的检索。首先构建了一个均衡的BCC检索式数据集，利用大语言模型为BCC检索式生成了自然语言描述。随后，微调大语言模型进使其能够支持自然语言到BCC检索式的转换。实验结果证明了TextToBCC模型的优异性能。这一成果不仅降低了BCC语料库的使用难度，而且有助于促进其在更广泛领域的传播和应用，为语言学研究和自然语言处理实践带来便利。

BCC检索式

- BCC检索式示例

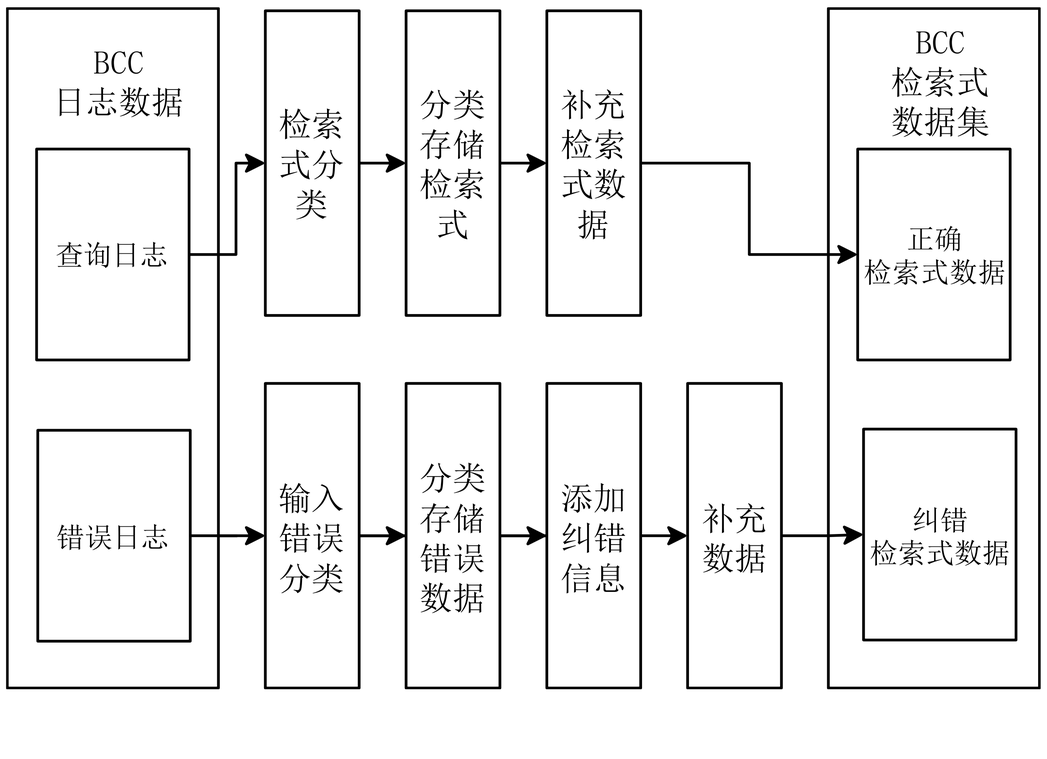
BCC检索式	中文表达
高大的n	高大的+名词
v了一v	动词+了一+动词
见*面	见后面离合出现面
洗.澡	洗后面隔一个字后接澡
../v	二字动词
爱(v)不(v){$1=$2}	爱+动词+爱+动词，并且两个动词相同
(nr) 说 m q {len($1)>1; begin($1)=[老]}	人名加说加数词再加量词，人名长度大于1并且以“老”字开头

- BCC检索式数据分布

类型	数量	比例
字符串	14767	41.18%
词性标签	8716	24.30%
通配符	836	2.33%
集合符	8065	22.49%
属性约束	561	1.56%
限制条件	2918	8.14%

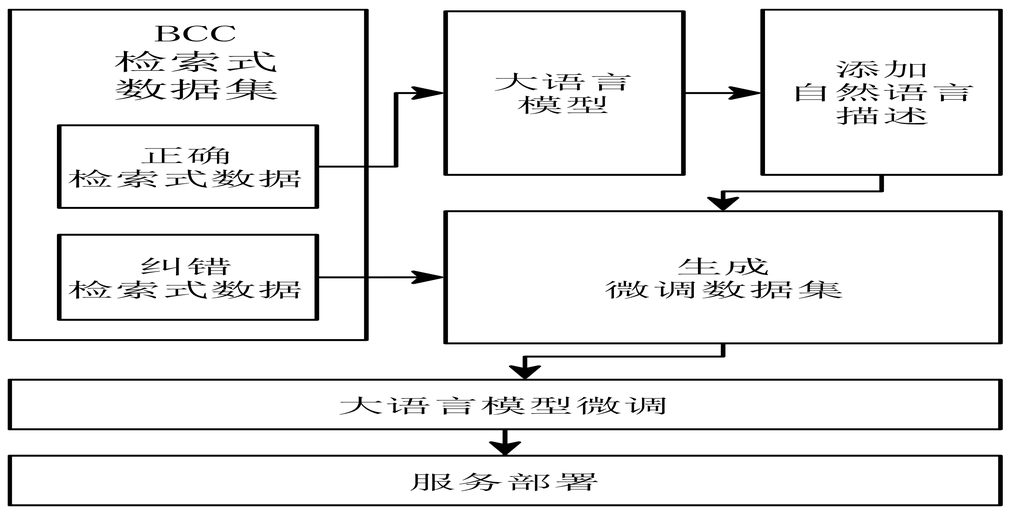
数据处理

- 抽取BCC检索日志数据和错误日志数据，构造检索式数据集。



TextToBCC模型

- 通过在线大语言模型生成BCC检索式的自然语言描述，生成微调数据集对本地大语言模型微调，得到TextToBCC模型。



- 大语言模型生成自然语言描述示例

检索式	通义千问	ChatGLM
./d高，支持率 ./d	单音节副词后接高，逗号，支持率，再接单音节副词	单音节副词后接高，并且高后面是逗号和空格，接着是支持率，支持率后面再接一个单音节副词
投票率 ./[d v]高	投票率后接单音节副词或动词，紧接着是高	投票率后面隔一个字接一个副词或动词，再接高
，~灿烂	句子中以逗号分隔，其后紧跟任何词性的一个词，最后是灿烂	逗号后面接一个词，再接灿烂
(v) 当 代 (vn){len($1)=2; end($2)=[高 妈]}	长度为2的动词后接当代，再接以高或妈结尾的名动词	动词后面接当代，当代后面接名动词，该动词是双音节并且名动词以高、妈结尾

实验

- 准确率测试

	Qwen-7B		ChatGLM3-6B	
	EM	EX	EM	EX
字符串	88.20%	89.47%	86.04%	86.42%
词性标签	72.17%	84.92%	67.63%	80.49%
通配符	73.94%	79.34%	71.83%	78.40%
限制条件	60.96%	74.07%	57.25%	70.68%
集合符	65.85%	72.70%	63.41%	69.69%
属性约束	56.29%	66.33%	55.27%	66.16%
平均	69.57%	77.81%	66.91%	75.31%

结论

TextToBCC为BCC语料库增添了自然语言查询的功能。通过构建BCC检索式微调数据集，并对大语言模型进行精细微调，顺利实现了自然语言向BCC检索式的转换。贡献如下：

1）构建了一个全面而均衡的BCC检索式数据集，为模型的训练提供了坚实的基础。

2）提出了BCC检索式与自然语言的联合训练方法，使得TextToBCC模型能够同时支持检索式和自然语言查询，显著提升用户体验。

3）引入纠错数据，使模型具备了BCC检索式的纠错功能，提高了模型的实用性。

https://github.com/paineliu/TextToBCC

Email: liutingchao@hotmail.com