arXiv:2304.05611v1 [cs.IR] 12 Apr 2023

FALQU：查找法律问题的答案

Behrooz Mansouri
南缅因大学，美国缅因州波特兰
behrooz.mansouri@maine.edu

Ricardo Campos
托马尔理工学院 Ci2—智慧城市研究中心 / INESC TEC，葡萄牙托马尔
ricardo.campos@ipt.pt

摘要

本文介绍了一个新的法律信息检索测试集合——FALQU：法律问题答案检索。该集合中的问题和答案取自 Law Stack Exchange（LawSE），这是一个面向法律专业人士及其他具有法律经验者的问答网站。与 Stack Overflow 类似，Law Stack Exchange 涵盖版权、知识产权和刑法等不同主题的各种问题，因此是构建数据集的有趣来源。问题也不限于单一国家。不同国籍的用户经常会询问不同国家的法律及相关专业领域。因此，FALQU 中的问题反映了现实用户的信息需求，有助于避免由实验室生成问题。另一方面，答案由该领域的专家提供。据我们所知，FALQU 是首个使用 LawSE 的测试集合，其问题比标准法律资格考试和司法考试中的问题更加多样。它包含 9,880 个法律问题和 34,145 条答案。除了这个新的测试集合，我们还提供了不同的基线系统，包括 TF-IDF 和 BM25 等传统信息检索模型，以及深度神经网络搜索模型。BM25 模型取得了最高的效果。

1 FALQU 测试集

尽管法律信息检索是一个新兴的研究领域，但随着一些法律问答数据集的涌现，它已成为研究工作的重点。最著名的当属 COLIEE-2015 [1], , 该数据集采用日本司法考试题目；以及 JEC-QA [2], , 该数据集采用中国国家司法考试的题目。尽管这些举措相继出现，现有数据集在问题类型和涉及领域方面仍缺乏多样性。Stack Exchange 等平台通过为不同领域提供社区问答网络，证明了这是一种解决该问题的有效方案。多年来，此类网络已应用于代码摘要 [3] 和数学信息检索 [4] 等任务。然而，尽管社区问答网站颇有用处，它们却从未用于法律信息检索。

在本研究中，我们使用 Law Stack Exchange[1] (LawSE) 构建了一个新的法律信息检索测试集合。法律信息检索通常被理解为寻找法律问题答案的过程；而在 LawSE 中，目标是找到单个答案。这与常见的信息检索任务有所不同：用户通常希望检索最相关的文档，而不是某一个特定文档。这类任务可以正式定义如下：给定一个由标题和正文表示的法律问题，信息检索模型应能够在参考数据集中的所有答案（帖子）里找到（搜索）并检索出相关答案（即提问者认定为相关的答案）。为构建测试集合，我们使用了从互联网档案馆获取的 LawSE 2022年10月8日快照。[2] 该快照包含 24,187 个法律问题，其中 10,129 个有采纳答案（即提问者认定为相关的答案）。为从数据集中剔除重复问题，我们利用了现有的重复链接功能。重复链接指向已发布的相同或几乎相同的问题。经过这一筛选过程，我们最终得到一个包含 9,880 个带有采纳答案的问题集合。

[1]https://law.stackexchange.com/ [2]https://archive.org 所提及的资源集根据互联网档案馆的政策，仅供学术和研究用途。

```
...
<问 题 >
   <编号 >17243</ 编号>
   <标题>应不应该有一封租赁   信          。。。  </ 标题>
   <BODY>我正在签租约..... </BODY>
</ 问题 >
...
```

```
...
<DOC>
   <DOCNO>  12  </DOCNO>
   <TEXT>国际上,                    根据          to...  </文本>
</文档>
...
```

图 1：FALQU 测试集合文件中的示例问题与答案。

选择问题　　对于训练集和测试集，我们首先根据.....将总计 9,880 个问题划分为 10 个分组　　n 问题的得分。[3]

按分数分箱可以确保训练集和测试集中的问题质量相近。分箱后，我们从每个箱中随机抽取问题（使用固定随机种子以确保结果可复现），将 90% 的问题划入训练集（8892 个问题），其余 10% 划入测试集（988 个问题）。每个集合都包含一个以制表符分隔值（TSV）文件格式存储的 TREC 格式 QREL 文件，其中有四列：query-number 0 document-id relevance-score；其中，query-number 是问题 ID，document-id 是答案 ID，而 relevance-score 始终为 1。在我们的设置中，每个问题在 QREL 中只有一个答案。该答案被视为相关答案（相关性得分为 1），因为提问者将其采纳为采纳答案。QREL 文件中未列出的其他答案均可视为不相关。

生成训练问题和测试问题后，我们汇总了所有答案（来自针对 9880 个问题获取的所有帖子），共得到 34,145 个答案。

整理后的答案采用 TREC 格式提供，包含 <DOC>、<DOCNO>（即 LawSE 上实际答案（帖子）的 ID）以及 <TEXT> 标签。问题以 XML 格式提供，每个问题都带有 <ID> 标签（即 LawSE 上实际帖子的 ID），以及 <TITLE> 和 <BODY> 标签，其中包含 LawSE 上问题的实际标题和正文及其对应文本。图 1 展示了 FALQU 测试集合中的一个示例问题（图上半部分）和一个示例答案（下半部分）。FALQU 测试集合及其生成代码已在 GitHub 上公开，供研究使用。[4] 为便于使用，我们将训练和测试主题文件及其相关的 QREL 文件分开提供。

对数据集的简要分析显示，FALQU 中的问题涉及多种主题，从“如果恶意网站窃取了我的信用卡信息，会发生什么？”这样的简单问题，到需要推理和历史知识的复杂问题，例如“在现代法律中，格林童话《汉塞尔与格蕾特》中主人公的行为会如何解读？”。还有一些问题涉及特定国家的法律，例如“作为一名伊朗人，我可以签署独立承包商协议，并从伊朗远程为一家欧盟公司工作吗？”。请注意，所有问题和答案均为英文。

2 基线模型

我们提供了多种基线系统，包括 BM25 和 TF-IDF 等传统信息检索模型，以及两种基于 BERT 的模型，采用 Sentence-BERT [5]。对于传统检索模型，我们考虑 2 种查询方式：(1) 将问题标题作为查询；(2) 将从问题正文中提取的关键词作为词袋，并将问题标题作为查询 +。

为提取关键词，我们使用了 YAKE [6] 关键词提取算法。每个问- 题提取排名前 5 的关键词。然后，为计算问题与答案之间的相似度，我们采用了 TF.IDF 和 BM25 PyTerrier

[3]该分数是 LawSE 的一个特征，计算方式为所有用户（不特指提问者）给出的赞成票与反对票之差；在此数据快照中，分数范围为 -9 至 226。[4]https://github.com/AIIRLab/FALQU

表 1：基线模型在测试问题上的平均倒数排名（MRR）@1000 和精确率@1。

模型	MRR@1000	P@1
TF-IDF	0.352	0.274
BM25	0.349	0.270
TF-IDF (YAKE)	0.407	0.313
BM25（YAKE）	**0.414**	**0.323**
distilroberta	0.337	0.243
all-MiniLM-L12-v2	0.372	0.293
distilroberta（微调）	0.368	0.283
all-MiniLM-L12-v2（微调版）	0.363	0.276

表2：使用 BM25和YAKE 时，带有P@1=1 (+答案) 和 P@1=0 (-答案) 的示例问题。（显示问题标题）

问题	儿童安排令 非生物学亲属的居住安排
+ 答案	这意味着你有权作出安排，处理诸如安排孩子出行之类的事情.....
问题	租赁协议到期后的通知期
- 答案	在惩戒学生和员工方面，私立学校基本上可以为所欲为.....

[7] 实现模型。对于 Sentence-BERT，我们考虑了两个预训练模型：“all-distilroberta-v1”和“all-MiniLM-L12-v2”。我们使用训练集中的所有问题对这两个模型进行了微调。对于每个问题，正样本对由问题及其采纳答案组成，负样本对则由问题和集合中的一个随机答案（除采纳答案之外的任何其他答案）组成。我们训练了 100 个周期，并将训练数据按 90:10 的比例划分为训练集和验证集。最佳参数使两种损失函数的组合值最小：对比[8] 损失和多重负样本排序 [9] 损失。

我们使用两项指标来比较各系统的有效性：平均倒数排名（MRR@1000）和 P@1。根据 TREC 运行中检索前 1000 条结果的标准，我们选取前 1000 条结果。这两项指标适用于本任务，因为每个问题只有一个相关答案。最终结果以宏平均值的形式列于表 1。如表所示，使用 BM25 模型与 YAKE 时，MRR@1000（按照 TREC 任务标准）和 P@1 均达到最高，显著优于其他所有基线模型；唯一例外是使用 YAKE 的 TF-IDF 模型。此结论采用学生 t 检验，p 值为< 0.05。对于基于 BERT 的模型，微调模型可能会略微提升“distilroberta”模型的表现。不过，与传统信息检索模型相比，无论是经过微调的还是预训练的基于 BERT 的模型，效果都较差。表 2 展示了两个问题，其中 BM25+YAKE 检索到了相关和不相关答案。观察 P@1=0 的情况可以发现，基线模型能够检索到可能相关的答案，但这些答案并非采纳答案，或者是针对其他类似问题给出的答案。这表明对候选答案进行人工标注十分重要，这将留待未来工作完成。

3 结论

本文介绍并发布了 FALQU（法律问题答案检索）这一新的测试集合，其中包含 8892 个训练问题和 988 个测试问题，每个问题均附有一个相关答案。此外，我们针对这项任务开展了实验，测试了 TF-IDF、BM25 和 Sentence-BERT 等不同基线模型。为衡量效果，我们采用了两项指标：Precision@1 和 Mean Reciprocal Rank@1000。综合这两项指标，使用问题标题和问题正文关键词的 BM25 模型取得了最高效果。我们希望未来法律信息检索领域的研究人员能够使用 FALQU，并将这一测试集合扩展至检索以外的用途，例如法律问题问答——在这类任务中，答案是生成出来的，而非检索得到的。

致谢

RicardoCampos 由国家基金通过 FCT - 科学与技术基金会（葡萄牙科学与技术基金会）资助，参与 StorySense 项目，项目编号为 2022年.09312.PTDC。

参考文献

[1] Mi-YoungKim、RandyGoebel和S Ken。COLIEE-2015：法律问题问答评测。载于第九届法律信息学国际研讨会（JURISIN 2015），2015年。[2] Haoxi Zhong、Chaojun Xiao、Cunchao Tu、TianyangZhang、Zhiyuan Liu和 Maosong Sun。JEC-QA：法律领域问答数据集。载于 AAAI人工智能大会论文集，2020年。[3] Srinivasan Iyer、Ioannis Konstas、Alvin Cheung 和Luke Zettlemoyer。使用神经注意力模型总结源代码。载于第54届计算语言学协会年会论文集，2016年。[4] Behrooz Mansouri、Vít Novotn`y、Anurag Agarwal、Douglas WOard和Richard Zanibbi。ARQMath-3（2022）概述：CLEF实验室第三届数学问题答案检索活动。载于欧洲语言跨语言评测论坛国际会议。施普林格，2022年。[5] Nils Reimers 和Iryna Gurevych。Sentence-BERT：使用孪生BERT网络的句子嵌入。载于2019年自然语言处理经验方法会议和第九届自然语言处理国际联合会议（EMNLP-IJCNLP）论文集，2019年。[6] Ricardo Campos、Vítor Mangaravite、Arian Pasquali、AlípioJorge、CéliaNunes和Adam Jatowt。YAKE!：利用多种局部特征从单篇文档中提取关键词。《信息科学》，2020年。[7] Craig Macdonald 和 Nicola Tonellotto。使用 PyTerrier在信息检索中进行声明式实验。载于 ICTIR 2020 论文集，2020年。[8] Xuxing Liu、XiaoqinTang 和Shanxiong Chen。通过判别式学习相似度度量并应用于古文字识别。载于知识科学、工程与管理国际会议。施普林格，2021年。[9] Matthew Henderson、Rami Al-Rfou、Brian Strope、Yun-Hsuan Sung、László Lukács、Ruiqi Guo、Sanjiv Kumar、Balint Miklos 和 Ray Kurzweil。面向智能回复的高效自然语言回复建议。arXiv预印本 arXiv:1705.00652，2017年。