北京大学学报(自然科学版)
Acta Scientiarum Naturalium Universitatis Pekinensis
doi: 10.13209/j.0479-8023.2018.068

特定领域问答系统中基于语义检索的非事实型问题研究

仇瑜[1,2,3]　程力[1,2,3,†]　Daniyal Alghazzawi[4]

1. 中国科学院新疆理化技术研究所，乌鲁木齐 830011; 2. 中国科学院大学，北京 100049; 3. 新疆民族语音语言信息处理实验室，乌鲁木齐 830011; 4. 阿卜杜勒阿齐兹国王大学计算机和信息技术学院，吉达 21493;
† 通信作者, E-mail: chengli@ms.xjb.ac.cn

摘要　面向财税领域非事实型问题，提出基于语义检索的方法来抽取答案。首先使用领域知识库对问题及领域文档进行语义标注，引入语义相似度特征提高法规及案例的检索准确率；其次使用排序学习算法融合领域文本的多种特征对法规检索结果优化；最后使用法规特征对案例检索结果进行筛选，并从相似案例中抽取相应答案。在真实数据集上的测试结果表明，该方法在准确率和效率上比基准方法有显著提升。

关键词　问答系统；非事实型问题；领域知识库；语义检索；排序学习

Semantic Search on Non-Factoid Questions for Domain-Specific Question Answering Systems

QIU Yu[1,2,3], CHENG Li[1,2,3,†], Daniyal Alghazzawi[4]

1. Xinjiang Technical Institute of Physics & Chemistry, Chinese Academy of Sciences, Urumqi 830011; 2. University of Chinese Academy of Sciences, Beijing 100049; 3. Xinjiang Laboratory of Minority Speech and Language Information Processing, Urumqi 830011; 4. Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah 21493;
†Corresponding author, E-mail: chengli@ms.xjb.ac.cn

Abstract A semantic-based retrieval method was proposed to extract answer sentences from tax regulation and cases. Firstly, a domain knowledge base was employed to generate semantic annotations for questions, regulations and cases. Secondly, a filtering system was developed for the removal of irrelevant cases from answer candidates. In addition, a semantic similarity measurement method was employed for answer extraction. Finally, a rank model was proposed for the optimization of the retrieved results. In order to validate the proposed method, a series of experiments were performed on real-life dataset. Experiment results show noticeable improvement in accuracy and performance compared to the baseline methods.

Key words　question answering system; non-factoid question; domain knowledge base; semantic search; learning to rank

不同于传统的信息检索(information retrieval, IR)，问答系统(Question Answering System, QAS)允许用户输入自然语言问句，系统返回结果不再是相关文档或网页的列表，而是一个准确地答案。

中国科学院“西部之光”基金(2017-XBZG-BR-001)、国家千人计划项目(Y32H251201)和中科院新疆理化所所长基金(2015RC007)资助
收稿日期: 2018–04–15; 修回日期: 2018–08–10; 网络出版时间: 2018-08-20 14:29:38

关于问答系统的研究已取得一定的进展，但研究重点主要集中于事实型问题(factoid)的问答，对于非事实型问题(non-factoid)问答的研究相对较少。

事实型问题询问的是某一客观事实，答案通常为实体或短语[1]。非事实型问题没有相对固定的范围和查找句式，答案通常为句子或段落[2]。目前非事实型问答系统的研究热点主要面向开放领域，系统从已有的问题库中搜索相似问题，然后把该问题的最佳答案反馈给用户，问题库包括常见问题集(frequent asked question, FQA)和社区问答集(community question answering, CQA)等[2]。相关测评任务(如 TREC[1]、CLEF[2]、NTCIR[3])关注如何从海量的文本中找到包含答案的文本片段，主要方法是使用信息检索技术在 Web 文档抽取相关答案，这些方法虽然在测试任务中取得不错的效果，但是在特定领域实际问答系统中还不能得到较满意的结果。原因在于特定领域的非事实型问题更加复杂多样，需要更深层次的理解用户查询需求；特定领域的答案获取过程借助更多的领域相关资源和领域文本特征。如何有效的理解用户意图，利用丰富的领域资源来提高领域问答系统的可用性及实用性成为关键问题[3]。

本文重点研究财税领域非事实型问题(比如，对询问某个涉税行为的处理方式)，这类问题不能用简单的事实作为答案，而是需要更长的答案(句子或段落)。信息源(答案源)为财税法规及案例，财税法规给出问题解答的依据，相似的案例直接提供问题的准确答案。在实际问题的分析中，财税领域专家也需要依据法规条款给出问题的答案，因此在缺乏相似案例时，可以使用相关法规条款作为推荐答案。

相比于其他领域，财税领域非事实型问题的答案抽取还面临如下挑战: 1) 财税用语中的事实及概念与日常不同，存在查询不匹配问题，即答案句中可能不包含问句中的词汇，只是具有语义相关性; 2) 财税的文档种类和结构较为复杂，各类属性特征是影响查询结果的重要因素，只考虑文本内容的相似性很难找到正确答案。

针对上述问题，本文引入领域知识库对问题和领域文档进行自动标注，将文本中的实体或概念映射到知识库中，发现更多语义信息，利用句子之间的语义相似度提高法规及案例的检索准确率；采用排序学习算法融合领域文本的静态特征、属性特征及关联特征等，建立排序模型，对检索结果重新排序，优化检索结果。

1　相关研究

非事实型问答系统能够回答关于寻求意见、方式、原因、定义等问题，相比于事实型问答系统，非事实型问答系统可回答的问题范围更加广泛[4–5]。研究者探索使用各类资源作为答案获取的途径，如 Fukumoto[6]使用模式匹配的方法在常见问答集(FAQ)中对“how”、“why”及“definition”类型的问题进行答案抽取。Tran 等[7]使用排序学习的方法，在社区问答集(CQA)中查找相似问题并抽取相关答案。Savenkov[8]结合社区问答集相似问题中抽取的候选答案集以及网络中检索的候选答案段落，使用线性模型筛选正确答案。这些方法虽然在通用领域取得了不错的应用效果，但是检索目标和范围都受到一定的限制，并且对于特定领域来说，领域术语、句子长度过长及特殊文本结构等问题使得这些方法在特定领域并不适用[7]。

目前关于财税领域的问答系统研究较少，大多数仍采用基于规则或关键字检索的方法。在相似领域(如法律领域)有很多研究值得参考，Prolo 等人[9]提出了对葡萄牙司法文件的问答系统，基于句法及语义进行问句分析，然后使用本体及逻辑推理得出答案。但是该系统需要大量人工标注数据，而且问题范围有限。Monroy 等[10]使用法规条款间的关系来构建图模型，并使用领域词典对法规条款进行检索，系统给出相关的条款作为问题的答案。但该研究仅用人工构建的有限的问题示例进行测试，而没有应用在实际场景中。2014 年，在法律信息抽取及蕴涵竞赛(COLIEE)中，将法律问答作为测试任务，测试问题为司法考试中的是非型问题。任务分为两个阶段，首先从法律文档中检索相关文档，然后根据检索到的法律条款判断问题的答案。Kim 等[11]针对上述任务，使用基于 TF-IDF 及主题模型 LDA 进行法律检索，并使用排序学习算法对检索结果排序；使用句法及语义相似度比较问句与相关法律条文来确定最后答案。这种非事实型问答只限于描述比较规范的考试问题，能够直接使用法规条款作为答案。

本文对财税领域非事实型问答的问题研究种类更加广泛，不仅限于有限的问题类别；获取答

[1] https://trec.nist.gov/tracks.html
[2] http://www.clef-campaign.org/2002.htm
[3] http://research.nii.ac.jp/ntcir/ntcir-14/tasks.html

案的途径更加多样化，结合财税法规及案例；检索的目标内容复杂，财税文档结构种类更加多样。

2　研究方法

本文将研究任务定义为在领域文本集中检索可能包含答案的段落(句子)，形式化表示为：给定一个非事实型问题 Q，法规文档集为 $\{D_1, D_2, ..., D_m\}$，案例集为$\{C_1, C_2, ..., C_n\}$。答案抽取的过程分为两步：首先从法规库中检索到相关的条款集$T=\{t_1, t_2, ..., t_s\}$，然后在案例集中检索到相似案例集$C_s=\{C_i, ..., C_k\}$，并从相似案例中抽取可能包含答案的句子集$S=\{S_1, S_2, ..., S_l\}$返回给用户。问答示例如图 1 所示，其中Q_1为用户查询问题，T_1, T_2, T_3为相关法规条款，$C-Q_2$为案例中的问题描述，A为案例中的答案。

Q_1: 赵某将位于丰台区丰科路的写字楼对外出租，提供的合同标明月租金收入 28000 元(不含税)，应该如何缴税?
T_1: 个人非住房出租月租金收入在我市营业税起征点(含)以上的，按照 12%的综合征收率计征各项税费…
T_2: 个人出租非住房取得的所得按 20%的税率征收个人所得税；个人出租住房取得的所得减按 10%的税率征收个人所得税。
T_3: 个人出租非住房月租金收入(不含税)在 30000 元(含)以下的，按 7%综合征收率计征税款；个人出租非住房月租金收入(不含税)在 30000 元(含)以上的，按 12%综合征收率计征税款。
C-Q_2: 王某某于 2016 年 5 月与李某签订《房屋租赁合同》，约定，租赁商铺位于…租赁期内每月租金为 30000 元，需要交纳多少税款?
A: 因此王某出租的房屋是非住房且月租金收入在 30000 元以下，应按 7%综合征收率缴纳税款。

图 1 财税问答示例

Fig. 1　Examples of tax questions and answers

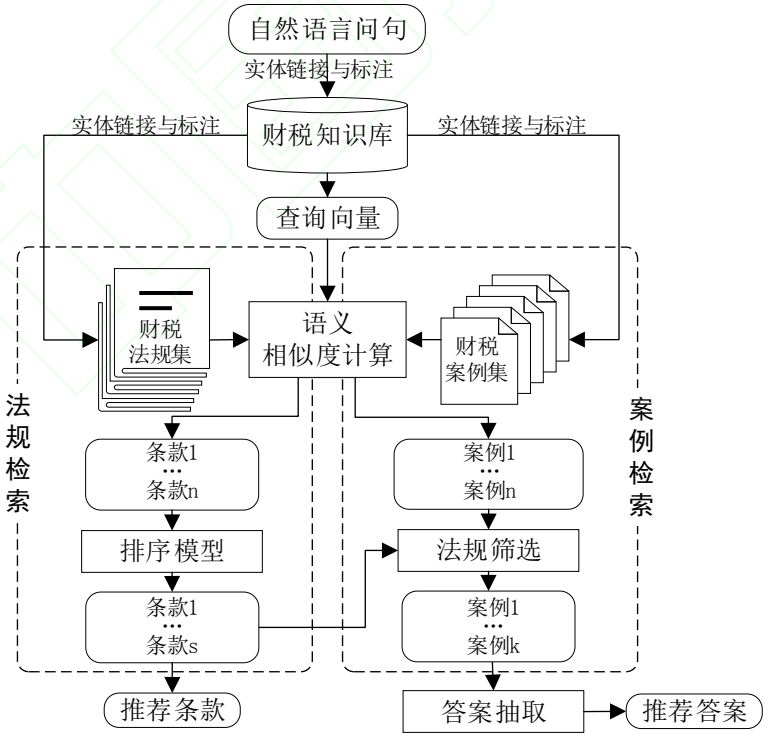

图 2 系统流程图

Fig. 2　System flow chart

问答系统的基本流程如图 2 所示。系统首先通过人机交互界面接受用户输入的自然语言问句，由问句析模块识别问句中的实体，经过实体链接与标注，对查询语句进行语义标注，将相关实体或概念链接到知识库，获得查询向量及语义标注信息。然后分别对法规和案例进行检索：在法规检索模块，使用语义相似度检索相关条款，并通过排序学习对检索结果排序，得到相关条款集；在案例集检索模块，使用语义相似度检索相似问题，并结合法规条款特征进行筛选，得到相似案例。最后从相似案例中抽取相关答案句推荐给用户，对于未在案例集中找到答案的问题，仅返回相关的法规条款作为参考。

2.1 财税知识库

财税知识库主要用来存储财税领域的概念、实体及关系等信息。首先我们使用半自动化的方法构建财税领域知识库，构建过程分为三个阶段[12]：1)领域专家根据可重用的相关本体及领域词典，构建财税领域的顶层本体；2)用规则+统计的方法，从基本法规中抽取出重要的概念及关系及实例，专家验证后加入顶层本体，构成初始的领域知识库；3)根据前两步得到的本体概念及部分实例，对领域实体进行训练，使用实体识别及标注方法，从财税文本中识别领域实体及关系，并映射到相应的概念类别。

目前知识库包含 1326 个概念、326 种关系及 235343 个实例。主要实体类别包括征税对象、文件、主体和地点等。

2.2 问句分析

对于用户输入的自然语言问句，首先进行实体链接与标注，使用知识库中的实例或概念对问句中的词进行标注，生成查询向量。

问句分析的总体流程如图 3 所示，首先对问句进行预处理，包括分词、词性标注(使用 ICTCALAS)和句法分析(使用 Stanford Parser)等。分词过程加入自定义规则(如 Hearst 模式[13])和用户词典(由知识库中的实体、概念及关系构成)来提高分词准确性。

传统的命名实体方法(如 Stanford NER)是由特定的标注语料训练的，可识别的实体类型有限，对于特定领域来说，由于缺乏相关训练语料而不能准确地识别出领域实体(如征税对象名“租金”没有被识别为实体)。为了尽可能多地抽取有用信息，本文根据词性标注结果及句法分析，将名词短语、动词、动词短语作为候选实体。

2.2.1 实体链接

实体链接对识别出的候选实体，使用 Levenshtein 距离[14]计算候选实体与知识库中实体之间的相似度，相似度大于一定阈值时，将候选实体链接到知识库中。

财税领域中普遍存在缩写的形式(如“个税”为“个人所得税”的缩写，“纳税人”为“纳税义务人”的缩写)，所以本文采用 Zhang 等[15]的启发式规则对缩略词进行扩展。此外由于财税领域实体歧义性相对较小，所以没有进行其他消歧处理。

2.2.2 实体标注

实体标注是针对没有链接到知识库中的候选实体进行标注。由于知识库固有的不完备性[16]，部分实体无法在知识库中找到对应项，本文通过监督学习的方法，对这些实体进行类别预测，将其映射到知识库的相应类别。该过程视为一个层次分类问题，支持多标签分类及部分深度标签(标签路径可以以非叶子节点结束)。本文参考 Yosef 等的方法[17]，对实体在知识库中的类别进行预测，具体步骤为：1)根据知识库中的概念结构定义层次标注集；2)采用远程监督的方法根据知识库中实体生成训练数据；3)在训练集中抽取实体的特征集，并训练分类器；4)根据训练好的分类器，对候选实体进行类别预测。

Yosef 等使用支持向量机(SVM)模型进行分类，存在训练数据不平衡问题及单个分类器的偏差(bias)问题，本文使用集成学习的方法进行改进。根据对分类器的差异度计算[18]，选择 SVM、逻辑回归及感知机 3 种分类器。使用 Bagging 方法进行集成，对于每个分类器的分类结果采用简单投票法(major voting)确定最终的实体类别[19]。本文实体分类使用的特征集如表 1 所示，实验证明分类预测效果优于单个分类器。

最后根据类别标注，将预测值大于阈值的实体映射到知识库中。例如，对于问句 Q_1，经过分析后将“赵某”标注为“个人”，知识库中存在实体“写字楼”为类“非住房”的实例，可以直接连接。

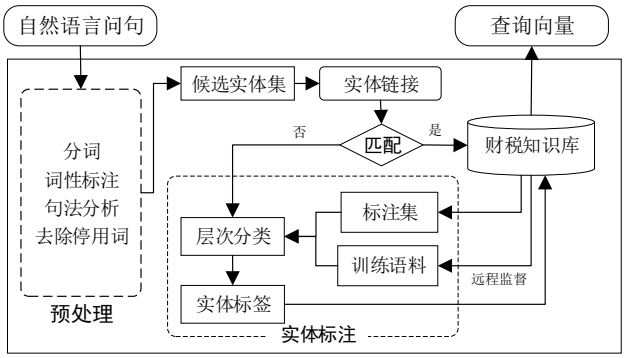

图 3 问句分析流程图

Fig. 3　Qeustion analysis flow chart

表 1 实体特征集

Table 1　Feature set of entities

特征	描述	示例
实体	实体本身的符号特征	粮食白酒
长度	实体的长度	4
上下文	窗口大小为 5 的上下文内容	“本月”“销售”“取得”“销售额”
聚类	实体进行词聚类的聚类中心 ID	78(饮料)
句法依存	句子中实体的依存关系信息	Patient_of: 销售
主题	实体所在文档的主题信息	“增值税”“消费税”
临近动词	句子中距离实体最近的动词	“销售”“取得”
词向量	训练得到的实体的词向量信息	“白酒”“生产”“价格”“葡萄酒”

2.3 法规条款检索

财税法规是问题答案或案例决策判断的重要依据，相关的法规条款具有重要的解释作用。由于问题是对实例层面的描述，法规则是对概念层面的描述，因此传统的基于关键词的检索方法很难取得较准确的结果。为了分析不同语义特征对检索结果的影响，同时提高排序学习的效率，本文中法规条款检索分为两个阶段：第一阶段使用多种语义相似度获取与用户问句相关法规条款集；第二阶段根据问句分析得到的查询向量，使用排序学习算法结合法规条款的多维度特征对查询到的条款进行重新排序。检索过程包括预处理、相似度处理和排序学习三个步骤。

2.3.1 预处理

根据法规结构，将法规文档分解为较短的段落(条款)，由于较短的条款可能不会包含查询词，所以本文将长度较短的条款与其父条款合并。然后进行预处理、实体连接与标注(方法同 2.2 节)。

2.3.2 相似度计算

本文使用 Fernando 等[20]关于释义识别(paraphrase identification)的方法，来计算问句与段落之间的文本相似度，使用相似度矩阵计算两个向量之间的相似度，相似度得分考虑了句子中每个词的相似度，计算公式如下：

$$sim(Q_e,T)=\frac{Q_eWT^T}{|Q_e||T|}, \tag{1}$$

其中，W 为语义相似度矩阵(包含任意两个词之间的相似度)，矩阵中 $w_{i,j}$ 表示问句中词 q_i 与条款中词 t_j 的相似度 $sim(q_i,t_j)$。传统的基于字符相似度的方法难以挖掘词汇之间的语义关系，本文使用词在知识库及语料库中的语义相关性计算 $sim(q_i,t_j)$。

目前基于知识库(如 wordnet)的语义相似度计算方法主要利用知识库中的层次分类关系[21]，如 Li 等[22]使用一种结合最短路径及深度的相似度计算方法(简称结构特征):

$$sim_{\text{path}}(a,b)=e^{-\alpha\cdot len(a,b)}\frac{e^{\beta\cdot h(a,b)}-e^{-\beta\cdot h(a,b)}}{e^{\beta\cdot h(a,b)}+e^{-\beta\cdot h(a,b)}}, \tag{2}$$

其中，$len(a,b)$ 为知识库中 a 和 b 的最短路径，$h(a,b)$ 为 a,b 最近公共节点到根节点的深度。α 和 β 为调节参数，经验值为 α=0.2 和 β=0.6。

此方法是针对知识库中层次分类关系结构的研究，大量的非分类关系没有被有效利用。本文将当前节点的所有非分类关系节点及关系本身作为节点的属性特征，然后根据特征集计算语义相似度(简称属性特征)，计算方法为

$$sim_f(a,b)=\frac{|F(a)\cap F(b)|}{|F(a)\cup F(b)|}, \tag{3}$$

其中 $F(a)$ 和 $F(b)$ 分别为 a,b 的具有非分类关系的实体、概念或关系集合。

本文借鉴 Lin 等[23]结合知识库结构和语料库，使用基于信息量(information content)的方法，通过比较共同祖先节点所包含的信息量来衡量相似度(简称信息量特征):

$$sim_{ic}(a,b)=\frac{2\times IC(lso(a,b))}{IC(a)+IC(b))}, \tag{4}$$

其中 $IC(a)$ 为 a 的信息量，$P(a)$ 是概念 a 在训练语料中出现的次数与训练语料总数比。$IC(a)=-\log(P(a))$，$lso(a,b)$ 为 a 和 b 最近的共同祖先节点。

我们使用 Word2Vec[24]训练的词向量(语料库法规及案例集)，发现词项之间的隐含语义信息，通过词项间的向量距离来计算相似度(简称词向量特征):

$$sim_{\text{emb}}(a,b)=\sum_{i=1}^{n}v_{ai}\bullet v_{bi}, \tag{5}$$

其中 n 表示训练向量的维度，v_a 和 v_b 分别表示 a 和 b 的词向量。

最后融合多种语义特征，对词项之间的语义距离进行计算:

$$w_{ij}=sim(q_i,t_j)=\delta sim_{path}(q_i,t_j)+\gamma sim_{ic}(q_i,t_j)+\lambda sim_f(q_i,t_j)+\mu sim_{emb}(q_i,t_j), \tag{6}$$

其中 $\delta,\gamma,\lambda,\mu$ 为相似度调节因子，且 $\delta+\gamma+\lambda+\mu$=1。

2.3.3 排序学习

语义检索利用词项之间的各种语义信息，丰富了查询结果，但同时也引入大量的噪音数据。查询结果仅依据语义相关性进行排序，排序标准单一，很多领域特征没有有效利用。由于财税法规结构复杂，法规条款之间具有引用关系，同时法规的各类属性(如发布时间、效力级别、使用范围等因素)对检索结果有重要影响，所以本文使用排序学习方法，融合法规多种的特征，对语义检索的结果重新排序。排序学习的流程如图 4 所示。

1)训练数据获取。

根据案例集和语义检索结果构建训练语料，首选从案例中抽取问题与条款的对应关系，构建 pair 对。然后用抽取的问题进行语义检索，将获取的结果再用案例集中的条款进行标注。最后根据标注语料，对排序模型进行训练。如根据问句查询得到 10 个条款，按顺序使用 t_i 进行表示，如果案例中含有 1, 3, 7 三条法规，优化的排序 1, 3, 7 应该排到前面，得到文档对的顺序关系。

2)排序学习算法。

排序学习是采用机器学习的方法训练模型来处理排序问题。本文使用 RankSVM[25]算法进行排序模型的训练。算法在训练集构造样本有序数据对，将排序问题转化为分类问题，使用 SVM 分类模型进行学习并求解。对给定查询 $x_u^{(i)}$ 和 $x_v^{(i)}$ 为相关条款，按评分大小得到偏序关系，$y_{u,v}^{(i)}$ 为相关性标签，$f(x)=w^Tx$ 为一个线性评分函数，w 为权重向量，C 为惩罚因子，损失函数如下所示:

$$\begin{aligned}&\min\frac{1}{2}\|w\|^2+C\sum_{i=1}^{n}\sum_{u,v:y_{u,v}^{(i)}}\xi_{u,v}^{(i)}\\&\text{满足}:w^T(x_u^{(i)}-x_v^{(i)})\geq 1-\xi_{u,v}^{(i)},\text{。}\\&if\ \ y_{u,v}^{(i)}=1,\xi_{u,v}^{(i)}\geq 0,i=1,...,n\end{aligned} \tag{7}$$

通过最小化损失函数训练，得到最优排序函数，对条款进行排序，获取排序列表。

3)特征选取。

在排序学习中，特征选取对于排序模型的预测结果有直接影响，本文参考 Liu[26]关于信息检索排序学习的特征总结，并考虑财税法规的属性特征，使用条款的静态特征、条款与问句的关联特征及问句本身的特征共 4 种特征作为特征集进行训练，如表 2 所示。

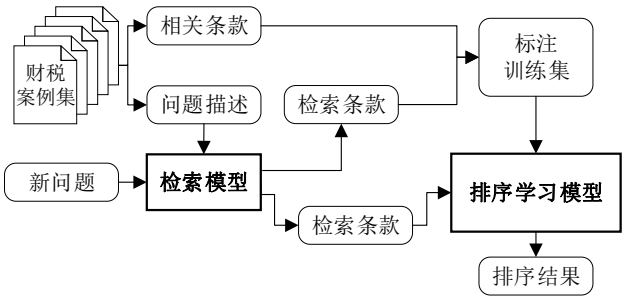

图 4 排序学习流程图

Fig. 4　Lerning to rank flow char

表 2 条款特征集

Table 2　Feature set of provisions

特征分类	特征维度
静态特征	TF、IDF、TF*IDF、包含问句词数、包含问句实体数和条款长度
属性特征	发布时间、效力级别、条款级别、适用范围、税种、行业、pagerank、引用数、被引用数和被查询次数
关联特征	BM25、LDA、LSI 和语义相似度
问句特征	问句长度、词数、TF、IDF、TF*IDF、实体数和问句类别

对于特征集中的非数值类型的特征(如类别属性特征)，我们使用 one-hot 方法处理[27]。由于每个特征的取值范围不同，需要对其归一化:

$$x_j^i = \frac{x_j^i}{\max(x_1^i, x_2^i, ..., x_m^i)}, \tag{8}$$

其中 j 为某一维特征，n 为特征维数，i 为数据集编号，m 为数据集数量。

法规库及案例库中具有时效属性特征，在对案例中的问句进行检索时，根据案例及法规条款的时效特征进行匹配，对失效条款进行过滤。排序学习训练完成后，对于新的问句查询，将语义检索结果属输入排序模型进行重新排序。最后选取排序结果 top N 作为条款检索最终结果。

2.4 案例检索与答案抽取

案例内容比较规范，包括问题描述、分析与结论。问题描述特征比较明显，可以使用规则的形式抽取，答案的抽取则需要更复杂的分析。答案获取的过程分为相似案例检索及答案抽取。

2.4.1 案例检索

案例的检索是根据用户输入的问题在案例集的问题描述中发现相似问题。由于案例是对某个具体涉税行为的描述，同类行为可能涉及不同的实体实例，因此关键字检索很难准确地找到相似案例(如图 1 中的 Q_1 和 C-Q_2)。本文使用语义相似度衡量新问题与案例问题的相似度，计算过程同 2.3 节，将相似度大于阈值的作为候选案例集。

为了更加精确的找到相似案例，本文使用排序学习对检索结果进行优化。由于缺少案例的标注，所以需要大量人工处理，本文仅利用法规引文对案例集进行筛选。筛选条件基于如下假设:问题的解答过程是法规条款的应用过程，相似的问题需要相似的法规条款进行解释。因此，具有相同参考法规的问题相似度更高。设 $T_q = \{t_{q1}, t_{q2}, ..., t_{qk}\}$ 为问题检索到条款集，$T_c = \{t_{c1}, t_{c2}, ..., t_{cl}\}$ 为案例中引用的条款集，使用如下公式计算条款集的相似性:

$$sim(q,c) = \frac{|T_q \cap T_c|}{|T_q \cup T_c|}。 \tag{9}$$

最后，将相似度小于阈值的案例从候选案例集中移除。

2.4.2 答案抽取

在案例中，答案的描述段没有明显的标识，无法用模板匹配的方法进行直接抽取，需要使用更复杂的方法识别和抽取答案句。目前答案句检索方法主要分为基于句子相似度和基于机器学习的方法[2,28]。由于缺少相关的训练语料，本文仅使用基于句子相似度的方法对答案句进行抽取。利用式(1)计算语义相似度，选取相似度最高的句子作为最终答案参考。

3 实验结果与分析

3.1 数据集

本文使用的数据集为半自动化方法构建的财税知识库以及从财税网站中爬取的公开法规集。

获取过程中可以通过相关描述得到发布时间、效力级别和有效性等结构化属性。案例集由专业服务人员提供。由于财税法规及案例具有较为规范的结构特点，对财税法规，通过机器学习及规则的方法进行分解，得到具有层次关系的条款集合及法规条款之间的引用关系；对财税案例，通过模式匹配的方法，抽取出问句描述及引用的相关法规条款信息，对其映射关系进行自动标注。最后获取财税法规集 22327 篇，条款集 415236 条，案例集 56261 篇，问题集 56261，问句与条款之间的映射关系 148372 个。

为了对比排序学习的效果，在法规检索与排序学习的测试集中，选取对应法规条款数大于 3 的问句作为法规条款检索及排序学习的测试语料，得到问句 33261 个，平均每个问句对应法规条款为 3.6 条。法规语义检索测试集使用自动获取的问题与条款的映射关系；排序学习测试集需要人工对排序结果进行标注。我们对检索结果的前 10 项顺序按相关性进行了人工标注，同一个任务分配给两个标注者，如果提交结果不同，则分配给权威专家进行标注，两个标注者标注的一致率约为 91%。

在案例检索与答案抽取的测试集中，随机选取 200 个案例进行人工标注。标注内容为相似案例及案例中的答案句，相似案例的标注一致率约为 93%，答案句标注一致率为 98%。

3.2 评估方法及结果

3.2.1 法规条款检索评估

测试使用案例中问题和相关法规条款的标注集。基准测试使用 BM25 及 TF-IDF 方法[26]，评价指标使用前 n 个结果的准确率(P@n)和平均查准率(mean average precision, MAP)。由于测试问句的相关条款大于 3 条，同时为了对比排序学习对检索结果排序的影响，本文选取 P@3 和 P@5 指标对检索结果进行分析。

为了进一步测试各语义特征影响，首先使用式 6(融合特征)进行测试，然后分别去掉单个语义特征，测试其对检索结果的影响(通过测试，当 $\delta,\gamma,\lambda,\mu$ 取值分别为 0.3,0.15,0.35,0.2 时效果最好)，结果如表 3 所示。

从表 3 看出，对于条款的检索，与基于关键字的方法相比，融合多种语义特征的方法在查询性能上均有明显的提高。主要原因是进行语义相似度计算时，考虑了更多字符不匹配的实体，如 Q_1 和 C-Q_2 中的“写字楼”和“商铺”在知识库中都属于“非住房”概念，通过结构特征分析可以得到较高的相似度值，而基于关键字的方法无法衡量这种关系。我们还随机选了 30 个问题检索结果，发现融合多种语义特征能够提高检索的召回率，这是由于增加问句的语义特征相当于对查询词进行了一定程度的扩展。表 4 显示基于知识库的各类语义特征与词向量特征都能在不同程度上提高检索的效果，其中去除结构语义特征，指标下降最明显，原因可能是实体的类别信息对检索过程影响较大(法规的内容多为概念性的描述)。此外对部分错误结果进行分析发现，很多错误是由实体链接与标注引起，错误的标注召回了更多不相关的内容。

3.2.2 排序学习评估

将实验一筛选后的案例集选取其中的 2/3，作为训练集，1/3 作为测试集。使用 SVMrank 工具[1](模型参数 C 设为 3)对语义检索结果进行排序，分析各类特征对排序结果的影响，分别去掉单个特征进行测试。评价指标使用平均查准率 MAP 及平均倒排序(mean reciprocal rank, MRR)用于对排序算法的评价，结果如表 4 所示。

表 3 法规条款检索测试结果

Table 3　Evaluation results of provision retrieval

方法	P@3	P@5	MAP
TF-IDF	0.5137	0.5368	0.2911
BM25	0.4953	0.5132	0.2836
融合特征	0.6752	0.6947	0.3418
–结构特征	0.6273	0.6441	0.3142
–属性特征	0.6431	0.6633	0.3265
–信息量特征	0.6672	0.6784	0.3219
–词向量特征	0.6457	0.6512	0.3143

表 4 法规条款排序学习测试结果

Table 4　Evaluation results of provision ranking

特征	MAP	MRR(3)	MRR(5)
全部特征	0.4136	0.6372	0.6128
–静态特征	0.3946	0.6147	0.5971
–属性特征	0.3627	0.5821	0.5714
–关联特征	0.4022	0.6217	0.6025
–问句特征	0.3946	0.6185	0.5933

[1] http://www.cs.cornell.edu/people/tj/svm_light/svm_rank.html

通过与表 3 对比，发现排序学习后 MAP 值提高了 21%，去除某类特征都会不同程度的降低 MAP 和 MRR 值，说明法规的领域相关特征对于优化查询结果是有效的。属性特征对排序结果指标下降最明显，证明法规条款属性特征对排序结果影响最大。通过结果分析我们发现一些规律，如中央法规一般比地方法规排序更靠前，新的法规比旧的法规排序更靠前，但是也有很多例外的情况。

3.2.3 案例检索和答案抽取评估

我们采用人工判别的方法对案例的检索进行评估。从案例集中随机抽取 200 个案例进行测试，使用案例问题描述检索相似案例，由领域专家评估返回的相似案例是否与检索问题相似，同时判断答案抽取的句子是否正确。案例检索评价指标使用 P@n、MAP 和 MAC。答案抽取指标使用 P@n、MRR 和召回率。

从表 5 可以看出，与关键字检索相比，语义特征对案例的检索效果同样有明显的提升，实验结果与法规条款的检索基本一致。使用相关条款对检索结果进行筛选，可以进一步提高准确率和 MAP 值，这也体现了之前的假说(即，法规条款是问题解答的依据，相似的问题具有需要相似法规条款进行解释)。对错误案例进行观察发现，多数错误的检索结果为含有数值型描述的问题，这是因为对于财税案例来说，数值的大小会有不同的处理方式(如金额和日期等)。

表 5 案例检索测试结果

Table 5 Evaluation results of case retrieval

方法	P@3	P@5	MAP
TF-IDF	0.5129	0.5363	0.2235
BM25	0.4718	0.5073	0.2131
融合特征	0.6874	0.7141	0.3175
–结构特征	0.6362	0.6537	0.2986
–属性特征	0.6724	0.6716	0.3034
–信息量特征	0.6658	0.6843	0.2912
–词向量特征	0.6536	0.6602	0.2841
条款筛选	0.7139	0.7354	0.3348

对于答案抽取的测试，首先选取检索的正确相似案例，对比基于关键词的句子相似性和基于语义的句子相似性计算方法，分别记为 S-关键字方法和 S-语义方法。为了评估法规检索和案例检索对于整个系统答案抽取作用，我们还测试了直接使用问句在案例文本中检索答案的方法，记为 P-关键字方法和 P-语义方法。关键字查询使用 TF-IDF 方法，测试结果如表 6 所示。

表 6 答案抽取测试结果

Table 6 Evaluation results of answer extraction

方法	P@3	P@5	MRR	召回率
S-关键字	0.6124	0.6431	0.2123	0.7240
S-语义	0.7436	0.7725	0.3517	0.8127
P-关键字	0.4316	0.4630	0.1822	0.4532
P-语义	0.5231	0.5832	0.2418	0.5641

在表 7 可以看出，语义特征能够有效提高答案句抽取的准确率、召回率和 MRR 值。与直接使用问句检索答案的方法相比，加入法规和案例检索过程对最终答案的获取效果提升显著。说明利用法规特征可以更准确的找到相似案例，证明了本文的方法的有效性。对实验结果分析发现，错误主要出现在较简短或包含多个句子的答案中，这类答案需要考虑更多的特征进行分析。

4 结语

本文主要对财税领域非事实型问题的答案抽取方法进行了研究。本文的主要贡献在于以下几个方面。

1)根据财税非事实型问答的领域资源特征，提出一种结合法规和案例资源的两阶段答案抽取方法，以法规条款为支撑，从案例中抽取相关答案，提高了答案获取的准确率。

2)针对用户问题表述不规范，查询匹配不准确的问题，使用领域知识库的多种语义特征对传统的语义检索算法进行改进，提高了法规和案例检索的效率。

3)针对领域文本结构复杂，领域特性难以有效利用的问题，使用排序学习算法，融合多种领域文本特征对检索结果进行排序，优化了法规检索的结果。

4)根据真实数据创建了财税领域非事实型问答测评集，并使用本文的方法进行测试，实验结果证明了该方法的有效性。对实验结果的分析与改进建议也为相关研究提供了重要的参考。

在将来的工作中，进行以下方面的研究: 1)改进实体链接与标注算法，提高语义标注准确率，以便更准确的获取文本的语义信息。2)进一步考

察和分析法规、案例及答案句的特征，研究特征选择方法[29]，对特征的重要性进行评估，选择最优特征集。3)对案例答案句进行人工标注，采用机器学习的方法，结合篇章结构、上下文、句子长度等多种特征进行训练，学习答案特征模型，进一步提高答案抽取的准确率。4)研究知识推理方法，发现隐含的语义信息，为答案抽取提供更多依据。

参考文献

[1] Surdeanu M, Ciaramita M, Zaragoza H. Learning to rank answers to non-factoid questions from web collections. Computational linguistics, 2011, 37(2): 351–383

[2] Yang L, Ai Q, Spina D, et al. Beyond factoid QA: effective methods for non-factoid answer sentence retrieval // European Conference on Information Retrieval. Padua, 2016: 115–128

[3] 王东升, 王卫民, 王石, 等. 面向限定领域问答系统的自然语言理解方法综述. 计算机科学, 2017, 44(8): 1–8

[4] Othman N, Faiz R. Question answering passage retrieval and re-ranking using n-grams and SVM. Computación y Sistemas, 2016, 20(3): 483–494

[5] 李舟军, 李水华. 基于 Web 的问答系统综述. 计算机科学, 2017(6): 1–7

[6] Fukumoto J. Question answering system for non-factoid type questions and automatic evaluation based on BE method // The 7th NTCIR Workshop. Tokyo, 2007: 441–447

[7] Tran O T, Ngo B X, Le Nguyen M, et al. Answering legal questions by mining reference information // JSAI International Symposium on Artificial Intelligence. Kanagawa, 2013: 214–229

[8] Savenkov D. Ranking answers and web passages for non-factoid question answering: emory university at TREC LiveQA // The Twenty-Fourth Text REtrieval Conference (TREC 2015). Gaithersburg, 2015: 1–8

[9] Prolo C, Quaresma P, Rodrigues I, et al. A question-answering system for portuguese // Knowledge and Reasoning for Answering Questions. Workshop associated with IJCAI05. Edinburgh, 2005: 45–48

[10] Monroy A, Calvo H, Gelbukh A. NLP for shallow question answering of legal documents using graphs // International Conference on Intelligent Text Processing and Computational Linguistics. Mexico City, 2009: 498–508

[11] Kim M Y, Xu Y, Goebel R. Legal question answering using ranking svm and syntactic/semantic similarity // JSAI International Symposium on Artificial Intelligence. Kanagawa, 2014: 244–258

[12] Qiu Y, Cheng L, Alghazzawi D. Towards a semi-automatic method for building Chinese tax domain ontology // 2017 13th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD, Guilin, 2017: 2530–2539

[13] Snow R, Jurafsky D, Ng A Y. Learning syntactic patterns for automatic hypernym discovery // Advances in neural information processing systems. Vancouver, 2005: 1297–1304

[14] Navarro G. A guided tour to approximate string matching. ACM computing surveys (CSUR), 2001, 33(1): 31–88

[15] Zhang T, Liu K, Zhao J. The NLPRIR entity linking system at TAC 2012 // Text Analysis Conference 2012. Gaithersburg, 2012: 1–5

[16] Gillick D, Lazic N, Ganchev K, et al. Context-dependent fine-grained entity type tagging [EB/OL]. (2014–12–03)[2016–08–01]. https://arxiv.org/abs/1412.1820

[17] Yosef M A, Bauer S, Hoffart J, et al. Hyena: hierarchical type classification for entity names // Proceedings of COLING 2012. Mumbai, 2012: 1361–1370

[18] Zhou Z H. Ensemble methods: foundations and algorithms. Taylor & Francis, 2012, 8(1): 77–79

[19] Breiman L. Bagging predictors. Machine learning, 1996, 24(2): 123–140

[20] Fernando S, Stevenson M. A semantic similarity approach to paraphrase detection // Proceedings of the 11th Annual Research Colloquium of the UK Special Interest Group for Computational Linguistics. Oxford, 2008: 45–52

[21] Rokach L. Ensemble-based classifiers. Artificial Intelligence Review, 2010, 33(1/2): 1–39

[22] Li Y, Bandar Z A, McLean D. An approach for measuring semantic similarity between words using multiple information sources. IEEE Transactions on knowledge and data engineering, 2003, 15(4): 871–882

[23] Lin D. An information-theoretic definition of similarity // The 15th International Conference on Machine Learning. Madison, 1998, 98: 296–304

[24] Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality // Advances in neural information processing systems. Lake Tahoe, 2013: 3111–3119

[25] Joachims T. Training linear SVMs in linear time // Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. Philadelphia, 2006: 217–226

[26] Liu T Y. Learning to rank for information retrieval. Foundations and Trends® in Information Retrieval, 2009, 3(3): 225–331

[27] Turian J, Ratinov L, Bengio Y. Word representations: a simple and general method for semi-supervised learning // Proceedings of the 48th annual meeting of the association for computational linguistics. Association for Computational Linguistics, Stroudsburg, 2010: 384–394

[28] Buscaldi D, Le Roux J, Flores J J G, et al. Lipn-core: Semantic text similarity using n-grams, wordnet, syntactic analysis, esa and information retrieval based features // Second Joint Conference on Lexical and Computational Semantics. Atlanta, 2013: 162–168

[29] Chandrashekar, G., Sahin, F. A survey on feature selection methods. Computers & Electrical Engineering, 2014, 40(1):16–28