

基于自适应扩展机制的领域智能问答系统

乔奋凤, 朱欣娟, 高 岭

(西安工程大学 计算机科学学院, 陕西 西安 710600)

摘 要:针对目前智能问答系统采用单层网络模型理解用户意图,未能准确关注用户语句中的细节特征的问题,提出了一种基于关键词分离的双层网络模型用户意图识别方法。第一层使用双向长短时记忆网络和条件随机场模型对用户语句中的关键词及问题句式进行识别,第二层将识别出的关键词作为细节特征,采用融合注意力机制的双层双向长短时记忆网络进行问题类型的识别,两层识别的结果为用户意图。实验证明,该方法的准确率和召回率平均提升了6%。针对用户数据较少时智能问答系统仍要扩展的需求,提出基于自适应扩展的智能问答系统优化方法。该方法使用基于句法结构的层次聚类算法对未识别的用户问题进行聚类,定期更新问题类型库。实验证明,基于句法结构的层次聚类算法正确率可达76%。

关键词:关键词分离;智能问答系统;用户意图识别;自适应扩展;句法结构;层次聚类

中图分类号:TP391

文献标识码:A

文章编号:1673-629X(2021)12-0013-08

doi:10.3969/j.issn.1673-629X.2021.12.003

Domain Intelligent Q&A System Based on Adaptive Extension Mechanism

QIAO Fen-feng, ZHU Xin-juan, GAO Ling

(School of Computer Science, Xi'an Polytechnic University, Xi'an 710600, China)

Abstract: In order to solve the problem that the intelligent question answering system uses the single-layer network model to understand the user's intention, but fails to pay attention to the details of the user's statements, we propose a hierarchical user intention recognition approach based on keyword separation. In the first layer, a bidirectional long short memory network and conditional random field model are used to pick the keywords out and label the sentence patterns. In the second layer, the previously selected keywords are used as input detail features, and a two-layer bidirectional long short memory network with attention mechanism is used to identify the question types. Experiment shows that the accuracy and recall of the proposed approach are improved by 6% on average. According to the needs that intelligent question answering system still needs to be expanded when the user data is small, an optimization method of intelligent question answering system based on adaptive expansion is proposed. The unrecognized user questions can be clustered by using the hierarchical clustering algorithm based on syntactic structure, and the question type library can be updated regularly. Experiments show that the accuracy of the hierarchical clustering algorithm reaches 76%.

Key words: keywords separation; intelligent question answering system; user intention identification; adaptive extension; syntactic structure; hierarchical clustering

0 引 言

智能问答系统的构建方法主要分为三种:基于模式模板匹配用户问题句式的方法、基于相似度匹配问题库的方法和基于知识图谱实时检索的方法。Karpagam 等人^[1]使用模式模板的方法对用户提交的查询进行问题类型的识别,依赖于早期构建的模式模板的完备性。Cai 等人^[2]使用问题对匹配的方法来实

现限定域内的中文智能问答。该类方法依赖于 FAQ 库的建立,不能根据用户需求实时更新 FAQ 库。知识图谱通过“实体-关系-实体”的形式来保存数据,可以有效地组织和表示知识,从而使知识本身得到有效利用^[3]。因此,基于知识图谱的智能问答系统成为智能问答系统构建的主流。

基于知识图谱的智能问答系统通常由用户意图识

收稿日期:2021-01-22

修回日期:2021-05-24

基金项目:国家重点研发计划项目(2019YFC1521405);陕西省重点研发计划(2019ZDLSF07-01);2020年西安工程大学研究生创新基金重点项目(chx2020027)

作者简介:乔奋凤(1995-),女,硕士研究生,研究方向为自然语言处理;通信作者:朱欣娟,教授,博士,研究方向为智能信息处理、三维虚拟展示;高 岭,教授,博士,研究方向为智能信息处理。

别、数据处理和答案检索^[4]三个模块构成。在用户意图识别阶段,如何分析用户问题语句,获取用户意图,最常用的方法是将用户问题语句进行分类,把分类结果看作用户意图。

基于知识图谱的智能问答系统目前仍处在起步阶段,进一步的发展需要从浅层语义理解逐步过渡到深层语义理解^[5]。要实现深层语义理解,重点应该关注用户意图识别,也就是问题分类。在用户意图识别方面,各类学者做了大量研究工作。Lim 等人^[6]针对文本数据以单词表示的特点,提出了基于语义张量空间模型的神经网络结构,从而对问题语句进行建模。Qiao 等人^[7]使用基于卷积神经网络(convolutional neural networks, CNN)的方法自动将句子归类。Xu 等人^[8]提出一种改进的基于 CNN 的跳格法的中文文本分类方法,并证明该方法比基于 CNN 的单热方法具有更高的性能。余本功等人^[9]使用多层级注意力卷积长短时记忆模型(multi-level attention convolution LSTM neural network, MAC-LSTM),它结合卷积神经网络与长短时记忆模型,并行提取词汇级特征,以此对用户问题语句进行建模,并实现问题分类,实验证明使用 MAC-LSTM 对用户问题类型进行识别,准确率远远大于使用传统深度网络模型进行用户问题类型识别。Wu 等人^[10]将注意力机制应用于文本分类中,验证了注意力机制的有效性。

目前智能问答系统普遍使用单层网络模型分析用户语句,由于单层网络模型识别能力的限制,造成对问题中的细节特征提取不足。如何充分挖掘用户问题中的语义信息,进行深层次用户意图识别,提高问题的分类精度是智能问答系统亟待解决的问题之一。同时,没有一个智能问答系统是一成不变的。随着用户需求的变化和知识的更新,智能问答系统中原先设定的问题类型和对应的知识也该随之迭代更新。常规的方法是使用机器学习、深度学习对网络上的知识进行提取,

以此来丰富系统知识。此方法依赖大量的用户数据,但是在系统搭建之初是缺乏大量用户数据的。因此,如何对系统中现有的问题类型和知识进行动态更新也是智能问答系统有待解决的问题之一。

针对以上智能问答系统构建存在的问题,笔者做了如下工作:

(1)使用双层网络模型进行用户意图的识别,重点关注了用户问题语句中的关键词,把它作为用户问题语句中的细节特征,作为输入信息添加到用户问题类型的识别模型中,实现了智能问答系统的深层次用户意图识别。

(2)为了使智能问答系统拥有更多的智能,该文提出了一种自适应扩展机制,利用人工和机器自动算法相结合的方式问题进行类型知识库的更新,实现了智能问答系统的扩展。

(3)在步骤(1)和(2)的基础上设计了针对学科课程问答的智能问答系统,该系统包括三个模块:意图识别、答案检索、数据更新。意图识别模块使用(1)中的基于关键词分离的用户意图识别技术;答案检索模块使用基于用户典型案例库的方法在知识图谱中进行答案的检索;数据更新模块使用(2)中的基于自适应扩展技术的智能问答系统优化方法。

1 智能问答系统知识库构建

领域智能问答系统依赖特定领域知识,该文以学科课程领域智能问答系统为例展开研究。智能问答系统后台数据库的构建包括两个内容:用户问题案例库的构建和学科课程领域知识图谱的构建。

1.1 用户问题案例库的构建

根据用户问题语句中的已知条件、待求答案、答案检索方式,将用户常见问题进行分类,分别对应每类用户意图,整理为用户问题案例库。构建好的典型用户问题案例示例如表 1 所示。

表 1 典型用户问题案例示例

用户问题	已知条件	待求答案	问题分析	问题类型
首先知道哪方面的知识会对我选修 A 课有帮助?	A 属于“课程”	知识点 B	如果先学习 B 知识,对学习 A 课程有帮助,那么表明 A 课程可以检验 B 知识的学习情况	课验知识
首先学习哪方面的知识会对我学习 A 知识有帮助?	A 属于“知识点”	知识点 B	首先学习 B 知识,对后续学习 A 知识有促进作用,说明 B 知识是 A 知识的先验条件	先验知识

当对用户问题在知识图谱中进行答案检索时,首先针对该问题类型的答案检索预设检索规则。而针对知识图谱来说,该检索规则应该和实体关系进行一一对应。但实际上,为了避免知识图谱存储的冗余,各种实体关系之间是可以进行推理的,所以最终呈现的基本实体关系边事实上与该检索规则不是一一对应的关系。因此,需要对问题类型进行一定的整理,以便和知

识图谱的基本实体关系进行一一对应。将问题类型定义如下:

定义 1:问题类型。按照不同答案检索规则分为三类,分别为“属性类问题”、“推理类问题”和“关系类问题”。对应知识图谱的实体和属性,检索规则为对属性进行检索的问题类型称为“属性类问题”;检索规则可以和知识图谱中的基本实体关系进行一一对应

的问题类型称为“关系类问题”;检索规则和知识图谱中的基本实体关系不能达到一一对应,但是可以通过知识图谱的基本实体关系推理得到,称这类型问题为“推理类问题”。

这三种问题类型对应的知识图谱检索规则如表2所示。

表2 三类问题类别的检索路径规则

问题类别	检索规则
属性类问题	首先找到问题中的中心起点类别实体,然后找到该实体所对应的属性,查找属性值
推理类问题	首先设定推理规则,根据推理规则在知识库中进行推理;再聚焦中心起点类别实体,顺着推理后的路径,找到检索终点
关系类问题	直接沿着路径进行检索

1.2 领域知识图谱的构建

领域知识图谱构建的基础是构建本体库^[11]。结合用户问题数据的特征设计本体模型^[12],设计好的教育技术领域学科课程知识图谱本体模型如图1所示。

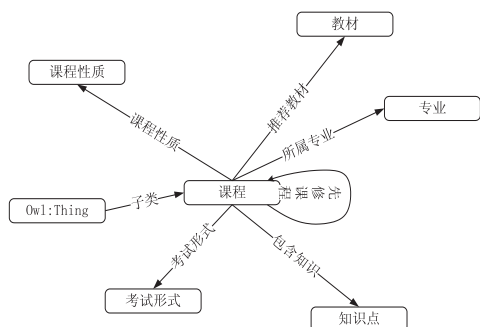


图1 学科课程知识图谱本体模型

文中学科课程知识图谱的初始数据来源为课程教学大纲,后续根据网络爬虫爬取维基百科的同类型词条对知识图谱进行数据扩充。

2 基于关键词分离的用户意图识别及基于典型用户问题推理的答案检索

2.1 基于关键词分离的用户意图识别

基于关键词的用户意图识别也就是把用户意图分为两层,第一层是用户问题语句中的关键词,第二层是用户问题语句所属的问题类型。

分析发现,用户问题中的关键词很大程度上影响了用户问题分类结果,也影响了用户意图判断,例如:

句子1:计算机网络管理这本书是哪门课上学的?

句子2:二叉树这个知识点是在哪门课上学的?

在这两个句子中,语法结构类似,所包含的词汇大体相同,仅有关键词不同,导致问题类型和检索路径也不同,具体如下:

(1)关键词类型不同:句子1中的关键词为“计算机网络管理”,属于“教材”实体类别;句子2中的关键词为“二叉树”,属于“知识点”实体类别。

(2)用户问题类型不同:句子1所属的问题类型为“教材课程”,句子2所属的问题类型为“知识课程”。

(3)答案检索路径不同:对用户答案检索时,句子1的答案检索规则为从“教材”出发,沿着“教材课程”边进行检索;句子2检索规则为从“知识点”出发,沿着“知识课程”边进行检索。

综合以上分析,用户问题中的关键词对用户意图识别有重要作用。常规的用户意图识别,如文献[7-9],缺乏对这种细节特征的充分利用,对用户语句整体进行问题类型提取,对问题类别识别率造成影响。

基于关键词分离的用户意图识别由两个子模型组成,分别为用户问题关键词识别模型和用户问题类型识别模型,总体结构设计如图2所示。

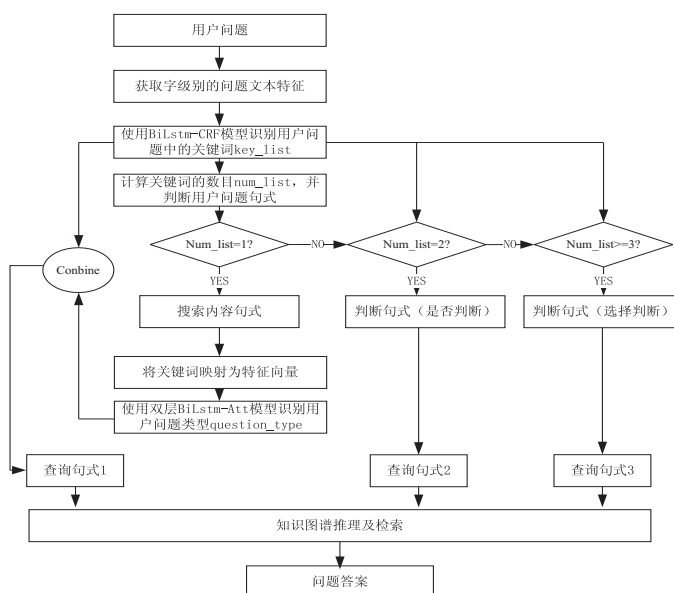


图2 基于关键词分离的用户意图识别模型结构

图 2 首先对用户意图进行初步提取,提取结果为用户语句中的关键词及关键词类别,将提取结果作为用户的第一层意图,用户第一层意图识别模型使用的是双向长短时记忆网络和条件随机场模型。

识别到关键词后,为了使识别到的用户意图更充分,也为了后期更方便地检索答案,根据关键词的数目对问题句式进行区分。对问题句式做如下定义:

定义 2:问题句式。根据问题语句中关键词数目的不同,将问题句式分为三种,具体如下:

If 关键词数目 ≥ 3 :

then 问题句式为“判断句式(选择判断)”;

If 关键词数目 = 2:

then 问题句式为“判断句式(是否判断)”;

If 关键词数目 = 1:

then 问题句式为“搜索内容句式”。

例如:

“判断句式(是否判断)”,如“《可视化程序设计》是选修课吗”,关键词为“可视化程序设计”和“选修课”,数目为 2 个;

“判断句式(选择判断)”,如“《地理信息系统》是

网络工程专业开设的还是软件工程专业开设的”,关键词为“地理信息系统”、“网络工程”、“软件工程”,数目为 3 个;

“搜索内容句式”,如“首先学习哪方面的知识会对我学习二叉树知识有帮助”,关键词为“二叉树”,数目为 1 个。

在用户问题类型识别模型中,首先使用依存句法分析工具^[13]对用户语句进行分析,得到分词结果和依存句法分析结果,然后融合关键词、分词、词性和依存句法分析结果,综合作为用户第二层意图识别模型的初始特征进行特征提取。用户第二层意图识别模型使用的是融合了注意力机制的双层双向长短时记忆网络(double-layer BiLSTM-att),双向长短时记忆网络(bidirectional long short term memory network, BiLSTM)可以有效利用文本的前后长距离特征^[2],使用双层特征提取可以使得特征提取更充分。采用注意力机制(attention)可以通过计算特征与结果的相似度差值对特征的权重进行调整^[2]。添加了关键词特征的用户问题类型识别模型结构如图 3 所示。

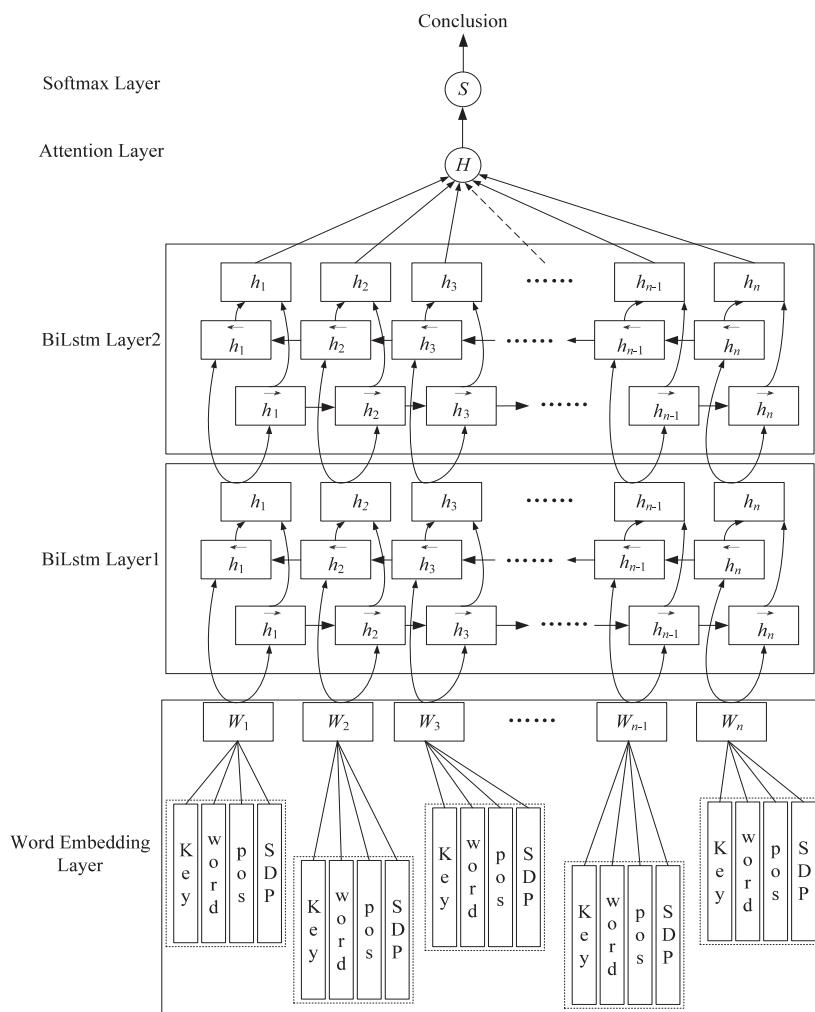


图 3 添加关键字特征的用户问题类型识别模型结构

图 3 中, key 表示关键词特征, word 表示分词特征, pos 表示词性特征, SDP 表示依存句法特征, W_1, W_2, \dots, W_n 表示经过特征映射后的词向量; $\vec{h}_1, \vec{h}_2, \dots, \vec{h}_n$ 表示双向长短时记忆网络的前向神经元; $\overleftarrow{h}_1, \overleftarrow{h}_2, \dots, \overleftarrow{h}_n$ 表示双向长短时记忆网络的后向神经元; h_1, h_2, \dots, h_n 表示双向长短时记忆网络综合前向和后向神经元的结果; H 表示注意力层; S 表示 softmax 分类层; Conclusion 表示最终结果。

最后, 综合用户的两层意图, 共同组合为用户的最终意图。这样, 使用双层网络模型, 关键词特征就更充分地利用起来了。

2.2 基于典型用户问题推理的答案检索

针对三种问题句式, 构建的检索方案如表 3 所示。

表 3 三种问题句式的答案检索方案

问题句式	关键词数目	检索方案
判断句式 (是否判断)	2	分别以两个关键词作为起始搜索节点, 构建查询语句, 在知识图谱中查找另一个关键词。如果查找到关键词之间有路径, 那么返回“是”, 如果以任意一个关键词为节点, 都没有查找到另一个关键词, 那么返回“否”
判断句式 (选择判断)	大于 等于 3	分别以三个关键词作为起始搜索节点, 构建查询语句, 在知识图谱中查找另外两个关键词。将查找到的结果返回, 即为该问题的答案
搜索内容 句式	1	以用户问题的关键词作为起始搜索节点, 问题类型作为边, 构建查询语句, 进行知识图谱的检索。最后, 把检索得到的结果返回, 即为问题的答案

根据表 3 的答案检索方案, 以及用户问题案例库中的检索路径, 对用户意图进行检索, 可以检索到用户所询问的知识词汇, 将该词汇插入到设定好的回答语句槽中, 组合完成的语句直接作为答案进行输出。

另外, 当用户检索问题时, 首先使用基于同义词词林的词语相似度计算方法^[14]将关键词与实体词典中的实体作对照, 作为对不规范关键词描述的修正, 输出该实体对应问题的答案, 即为用户所提问题的答案。当用户对给出的答案仍不满意时, 可以选择对答案进行评分。

3 基于自适应扩展的智能问答系统优化方法

当系统输出答案时, 允许用户对答案进行评分, 系

统将接收评分, 并将其写入反馈数据库, 自适应扩展就在于定期对反馈数据库中的用户反馈进行处理。

定期对反馈数据库中的数据进行整理。用户反馈主要为两种: 一种是需要扩充用户问题案例库, 即用户提出新的类型问题。另一种是答案不在知识库中。对于前一类反馈, 首先使用基于句法结构的层次聚类算法进行问题的初步聚类; 然后人工对新问题类型进行命名, 更新用户问题案例库; 最后通过网爬数据及官方数据进行知识图谱实例更新。对于后一类反馈, 只需扩充数据来源, 进行知识图谱实例更新即可。

自适应扩展智能问答系统, 也就是自适应扩展用户问题案例库, 重点在于对用户反馈进行分析处理。考虑用户语句的特点, 使用基于句法结构的层次聚类算法^[15]。文献[15]表明在中文文本领域, 该算法能够准确地进行聚类分析, 并能得到很好的聚类效果。

3.1 文本的句法结构

经分析发现, 句法结构对文本含义的影响很大, 比如用户问题语句为:

- A. 计算机包含数据库吗?
- B. 网络包含 TCP/IP 吗?
- C. 我的计算机里面有数据库软件吗?

如果仅考虑文本之间的相似度, 由于句子 A 和句子 C 中都包含“计算机”和“数据库”, 那么 A 和 C 将被划分为一类, 但是考虑用户意图, 句子 A 是询问“计算机”这门课中是否包含“数据库”这个知识点; 句子 B 是询问“网络”这门课中是否包含“TCP/IP”这个知识点; 句子 C 是询问“计算机”这个设备中是否包含“数据库”软件。所以句子 A 和 B 属于同类型问题。根据分析发现句子 A 和句子 B 句法结构类似, 均为 a 包含 b 的句式, 可以先把这样的句式提取出来, 再进行下一步聚类。

3.2 层次聚类算法

对于问题文本来说, 在聚类前并不知道将要操作的数据有多少种类型, 故使用分层聚类算法, 预先不划定聚类簇数, 而是在操作中对聚类簇数进行更新。在聚类时, 计算待比较文本的相似度, 如果相似度在阈值内, 则进行更新; 如果任意两个句子的相似度都不在阈值内, 那么结束聚类操作。

对如下两个数据预处理后的句子 S_i 和 S_j :

$$S_i [(w_{1i}, w_{1i'}) p_{1i}, (w_{2i}, w_{2i'}) p_{2i}, \dots, (w_{mi}, w_{mi'}) p_{mi}]$$

$$S_j [(w_{1j}, w_{1j'}) p_{1j}, (w_{2j}, w_{2j'}) p_{2j}, \dots, (w_{nj}, w_{nj'}) p_{nj}]$$

其中, $w_{1i}, w_{1i'}, w_{2i}, w_{2i'}, \dots, w_{mi}, w_{mi'}, w_{1j}, w_{1j'}, w_{2j}, w_{2j'}, \dots, w_{nj}, w_{nj'}$ 表示经过句法结构分析后的词; $p_{1i}, p_{2i}, \dots, p_{mi}, p_{1j}, p_{2j}, \dots, p_{nj}$ 表示句法分析符号。

句子 S_i 和句子 S_j 的相似度 sim_{ij} 的计算公式为:

$$\text{sim}_{ij} = \frac{\sum_{k=1}^{\text{SCount}} \max \{ \text{match}(S_{ki}, S_j) \}}{\max \{ \text{SCount}_i, \text{SCount}_j \}} \quad (1)$$

$$\begin{aligned} \text{match}(S_{ki}, S_j) = & \max \{ \text{similarity}(w_{ki}, w_{lj}) + \\ & \text{similarity}(w_{ki}, w_{lj'}), \dots, \\ & \text{similarity}(w_{ki}, w_{nj}) + \\ & \text{similarity}(w_{ki}, w_{nj'}) \} \end{aligned} \quad (2)$$

其中, SCount_i 为句子 S_i 的有效搭配对数, SCount_j 为句子 S_j 的有效搭配对数, S_{ki} 为句子 S_i 的匹配对。 match 为匹配函数。 $\text{similarity}(w_i, w_j)$ 为相似函数, 结果通过计算同义词词林中 w_i 和 w_j 对应的义项编号的距离得到^[14], \max 为最大值函数。

在结合新簇后, 需要对簇中心点进行更新, 中心点的计算方法为: 计算簇中每个句子与其余句子的平均相似度, 平均相似度最大的句子为该簇的中心点。

3.3 基于句法结构的层次聚类算法

基于句法结构的文本聚类方法步骤如下:

第一步: 数据预处理。

(1) 提取关键词^[16], 避免关键词语义对文本分析的影响。处理方法为: 使用意图识别模型的第一层模型提取用户问题中的关键词, 并将关键词所代表的词语用符号表示, 不是关键词的内容继续用原来的词语表示。根据第 1 部分的描述, 关键词的种类有 6 种, 关键词及对应符号为: “课程” → C、“教材” → B、“课程性质” → Y、“考试形式” → T、“专业” → S、“知识点” → K。

(2) 提取文本的句法结构。

分析文本的句法结构, 使用中文依存句法分析工具^[13]进行句法结构分析, 并将句法结构分析的结果作为待比较文本的替代。

对于原句子: “数据通信原理考试形式是什么”, 经过预处理后的结果为“(C, ‘考试形式’), (‘考试形式’, ‘是’), (‘是’, ‘root’), (‘什么’, ‘是’)”。其中, root 是句法结构分析的根, 所有句法结构关系都由这个根发出, 这个根由文献[13]的句法结构分析工具自动生成。

第二步: 分层聚类。

对预处理后的两个句子进行分层聚类。

基于句法结构的分层聚类算法描述如下:

算法: 基于句法结构的分层聚类算法。

输入: 预处理后的文本序列 DataList ; 相似度阈值 T ;

输出: 聚类后的结果 conclusion 。

1: for sen in DataList ;

2: 计算相似度最大的两个文本($\text{sen1}, \text{sen2}$)的相似度

3: if $\text{sim} \geq T$;

4: $\text{conclusion.append}(\text{sen1}, \text{sen2})$ // 把这两个句子加入

聚类结果

5: 计算聚类簇的中心点 center

6: $\text{DataList.delete}(\text{sen1}, \text{sen2})$

7: $\text{DataList.append}(\text{center})$ // 用中心点替换原来的两个句子

8: else:

9: 聚类结束, 输出结果

进行基于句法结构的层次聚类后, 形成 C_1, C_2, \dots, C_m 共 m 种新的问题类型, 为了保证问答助手的准确性, 人工对聚类结果进行二次筛选, 并给新的问题类型命名, 将其写入用户问题案例库中。新类型问题的解决依赖大量知识图谱实例, 故采取网络爬虫和扩充数据来源的方式进行知识图谱实例更新。

4 实验与结果

在本节中, 使用该文提出的方法构建了一个基于自适应扩展的智能问答系统, 并在教育技术领域用户常用问题集上进行了测试。

4.1 实验数据

基于知识图谱的智能问答系统数据一方面依赖知识图谱的完备性, 另一方面依赖问答分类标准。对于教育技术领域的问题分类, 目前还没有公开的数据集, 而其他领域的数据集也不适用于该领域。为了解决这一问题, 利用某大学计算机学院的课程大纲构建了面向学科课程知识问答的知识图谱雏形, 并使用爬虫技术爬取了同类型实体词条进行扩充。

用户意图识别的数据来源有两个:

(1) 公开领域有各种类型的用户问题语句, 根据课程大纲数据在这些问题的基础上进行修改。

(2) 网爬各大教育平台及教务系统的学生问题。

教育技术专家制定教育技术领域的数据标注标准, 使用众包的方式对使用以上两种方式收集的用户语句进行标注。为了避免小样本数据造成的过拟合, 提高模型鲁棒性, 首先对样本中的名词使用同义词替换的方式进行数据增强, 最终生成针对教育技术领域面向学科课程的智能问答系统用户意图识别小型数据集 CCID(Chinese course intention dataset)。

4.2 度量标准

在意图识别的实验中, 使用准确率、召回率、F1 值作为意图识别结果的评价指标。其中, 准确率指在分类结果为某类别的样本中, 实际上属于该类别的样本个数占该样本个数的比值; 召回率指实验分类结果为某类别的样本个数占实际上应属于该类别的样本总数的比值; F1 值(f1-score)为准确率与召回率的算术平均数。

文本聚类的判断标准为准确率, 使用公式(3)、公式(4)来计算用户问题类型聚类结果的正确率 preci-

sion。

$$\text{precision} = \frac{C_1 + C_2 + \cdots + C_n}{n} \quad (3)$$

$$C_i = \frac{\text{Count}_i}{S_i} \quad (4)$$

其中, C_i 表示聚类结果中代表文本类型为 i 的聚类簇的聚类正确率; n 表示聚类簇数; S_i 表示聚类结果为文本类型 i 的聚类簇中文本的总个数; Count_i 表示聚类结果本该为文本类型 i 的文本个数。在以上公式中, i 的取值范围均为 $1 \sim n$ 。

4.3 实验和结果

进行两个实验来验证提出的自适应扩展模型。

第 1 个实验验证使用关键词分离的方法提高了用户意图识别的准确率。总体采用十次十折交叉验证的方式进行测试。本实验将文中方法与其他文献中的网络模型结果进行多次实验对比。表 4 为文中双层模型的超参数。实验得到的数据结果如表 5 所示。

表 4 模型参数设置

参数	说明	取值
Num-filters	神经元的个数	100
Embedding_dim	词向量维度	350
Epochs	模型迭代次数	100
Batch_size	批处理大小	100
Dropout	防止过拟合参数	0.15

表 5 文中方法与其他网络模型对问题类型识别的结果对比

方法	precision	recall	F1 值
CNN ^[7]	0.62	0.64	0.63
SVM+withKey ^[17]	0.65	0.67	0.66
跳格 CNN ^[8]	0.68	0.65	0.66
单层 BiLSTM ^[18]	0.69	0.72	0.70
MAC-LSTM ^[9]	0.73	0.74	0.73
文中方法 (withoutKey)	0.80	0.77	0.78
文中方法 (withKey)	0.87	0.82	0.84

从实验结果可以得出,文献[9]的 MAC-LSTM 使用的双层网络模型较传统的神经网络有明显的效果提升。而文中提出的模型优于 MAC-LSTM,具体表现为:在文中用户意图识别模型不添加关键词特征时,比 MAC-LSTM 的正确率提高了 7%,召回率提高了 3%,F1 值提高了 5%,原因是文中在问题类型识别模型中使用了双层双向长短时记忆网络,比普通的长短时记忆网络能获取更多的特征;在文中用户意图识别模型添加关键词特征时,比 MAC-LSTM 的正确率提高了 14%,召回率提高了 8%,F1 值提高了 11%。这是由于相比于 MAC-LSTM,文中提出的方法将文本中的

关键词作为特征。使用文中提出的用户意图识别模型,添加关键词特征比不添加关键词特征的准确率提高了 7%,召回率提高了 5%,F1 值提高了 6%。由此说明,添加关键词特征的双层意图识别模型是有效的。这是因为关键词特征在问题类型识别中占有很大的作用。比如对于“课验知识”和“先验知识”,在不添加关键词作为特征时,这两类类型易混淆,它们的含义都是已知一个类别,求这个类别的先驱知识。如果不考虑关键词特征,光从语义上来看,并不能确认某一个问题是课验知识还是先验知识。但是添加了关键词特征后,如果关键词类别为“知识点”,那么这个问题就被划分为“先验知识”,如果关键词类别为“课程”,那么这个问题就被划分为“课验知识”,由此可见,提前进行关键词的识别在一定程度上可以减轻问题类型识别的混淆度。

第 2 个实验是选择在哪个阈值下基于句法结构的分层聚类算法能得到聚类最好效果。在其他条件不变的情况下调整阈值 T ,并计算最终得到的聚类正确率,得到的实验结果如图 4 所示。图中横坐标为阈值,纵坐标表示在该阈值下聚类结果的正确率。

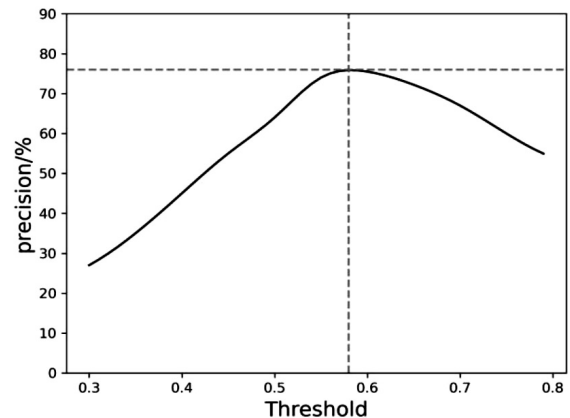


图 4 改变阈值对基于句法结构的分层聚类算法的正确率的影响对比

从图 4 中可以看到,当阈值 T 取值为 0.58 时,聚类的正确率达到了 76%,在这样的正确率下,可以在一定程度上更新智能问答系统的后台数据。

通过以上实验结果,可以得到结论:添加了自适应扩展机制后,智能问答助手有更好的用户满意度。究其原因,是因为深度学习方法需要大量用户数据,而智能问答系统搭建之初,用户数据较少,文中自适应扩展机制使用的是基于句法结构的层次聚类算法,这种方法依赖较少的数据,在一定程度上可以满足对智能问答系统数据更新的需求。另外,该自适应扩展机制使用定期更新的方式对用户反馈进行处理,可以动态满足用户的需求。同时,定时对后台数据进行更新,也有助于构建更可靠的智能问答系统。

5 结束语

该文提出一种基于自适应扩展机制的智能问答系统构建方法,并以学科课程领域知识问答为例。首先构建用户问题案例库和知识图谱;然后对于目前大多数智能问答系统使用单层模型不能充分利用用户语句中的细节特征、不能获取用户深层语义、从而不能更好地识别用户意图的问题,提出基于关键词分离的双层模型识别用户意图方法。实验表明,该方法可以有效提高用户问题意图的识别精度。针对系统构建之初用户数据量较少,智能问答系统需要进行后台优化的问题,首先使用基于句法结构的层次聚类算法对用户语句进行初步聚类,然后人工对聚类结果进行二次核查并更新用户问题案例库,达到智能问答系统自适应扩展的目的。提出的基于关键词分离的双层模型识别用户意图方法和自适应扩展机制,基于特定领域知识图谱,可以进行拓展应用。目前智能问答系统构建技术仍有很大的发展空间,未来的研究工作将更多着眼于寻求在自动聚类时,提高识别新问题类型的成功率的方法,以此进一步减少人工参与。

参考文献:

- [1] KARPAGAM K, SARADHA A. A framework for intelligent question answering system using semantic context-specific document clustering and Wordnet [J]. Sadhana, 2019, 44 (3): 62.
- [2] CAI L Q, WEI M, ZHOU S T, et al. Intelligent question answering in restricted domains using deep learning and question pair matching [J]. IEEE Access, 2020, 8: 32922-32934.
- [3] CHEN X, JIA S, XIANG Y. A review; knowledge reasoning over knowledge graph [J]. Expert Systems with Applications, 2019, 141: 112948.
- [4] MALIK N, SHARAN A, BISWAS P. Domain knowledge enriched framework for restricted domain question answering system [C]//IEEE international conference on computational intelligence & computing research. New York, USA: IEEE, 2014.
- [5] 王智悦, 于清, 王楠, 等. 基于知识图谱的智能问答研究综述 [J]. 计算机工程与应用, 2020, 56(23): 1-11.
- [6] LIM P. A tensor space model based deep neural net-work for automated text classification [J]. Database Research, 2018, 34(3): 3-13.
- [7] HUANG Qiao, XIA Xin, LO D, et al. Automating intention mining [J]. IEEE Transactions on Software Engineering, 2020, 46(10): 1098-1098.
- [8] XU W, HUANG H, ZHANG J, et al. CNN-based skip-gram method for improving classification accuracy of Chinese text [J]. KSII Transactions on Internet and Information Systems, 2019, 13(12): 6080-6096.
- [9] 余本功, 许庆堂, 张培行. 基于 MAC-LSTM 的问题分类研究 [J]. 计算机应用研究, 2020, 37(1): 40-43.
- [10] WU C, LUO G, GUO C, et al. An attention-based multi-Task model for named entity recognition and intent analysis of Chinese online medical questions [J]. Journal of Biomedical Informatics, 2020, 108: 103511.
- [11] LI X, WU Z, GOH M, et al. Ontological knowledge integration and sharing for collaborative product development [J]. International Journal of Computer Integrated Manufacturing, 2018, 31(3): 275-288.
- [12] KEJRIWAL M, SEQUEDA J, LOPEZ V. Knowledge graphs: construction, management and querying: editorial [J]. Semantic Web, 2019, 10(6): 1-2.
- [13] 哈尔滨工业大学社会计算与信息检索研究中心. 中文依存句法分析 [EB/OL]. (2013-01-16)/[2020-09-20]. <http://ir.hit.edu.cn/>.
- [14] 杨泉, 孙玉泉. 基于《同义词词林》深度的词义相似度计算研究 [J]. 计算机工程与应用, 2020, 56(17): 48-54.
- [15] 尹积栋, 谢茶花, 彭崧, 等. 基于句法结构分析的中文文本聚类方法研究 [J]. 计算机与数字工程, 2018, 46(5): 933-935.
- [16] QIAO Fenfeng, ZHU Xinjuan. Domain intelligent Q&A user intention recognition based on keyword separation [C]//2020 international conference on culture-oriented science & technology. New York, USA: IEEE, 2020: 224-229.
- [17] 汤铭. 基于领域知识库的校园智能问答系统关键技术研究 [D]. 南京: 东南大学, 2018.
- [18] 刘依红, 杨波, 孙宇宁, 等. 基于 BiLSTM 的婚姻法自然语言问答 [J]. 计算机工程与设计, 2019, 40(4): 1190-1195.