天津大学

本科生毕业论文

题目：基于知识图谱的问答系统的设计与实现

学　　院　智能与计算学部

专　　业　计算机科学与技术

年　　级　2018

姓　　名　刘煜堃

学　　号　3018216233

指导教师　饶国政

独创性声明

本人声明：所呈交的毕业设计（论文），是本人在指导教师指导下，进行研究工作所取得的成果。除文中已经注明引用的内容外，本毕业设计（论文）中不包含任何他人已经发表或撰写过的研究成果。对本毕业设计（论文）所涉及的研究工作做出贡献的其他个人和集体，均已在论文中作了明确的说明。本毕业设计（论文）原创性声明的法律责任由本人承担。

论文作者签名：

年　　月　　日

本人声明：本毕业设计（论文）是本人指导学生完成的研究成果，已经审阅过论文的全部内容。

论文指导教师签名：

年　　月　　日

摘　要

知识图谱可以很好的对领域知识进行建模与储存，但仍无法直接满足普通用户的使用需求，人们更习惯于通过问答方式来获取答案，因此亟需建立基于知识图谱的问答系统。

本文针对以上需求，首先针对基于 BERT 的经典问答方法在回答中文自然语言问句时准确率相对较低的问题，在经典的 BERT-BiLSTM-CRF 模型的基础上，进一步融入了注意力机制，得到了改进后的 BERT-BiLSTM-Att-CRF 模型。其次，本文研究了基于语义模板的问答方法，并根据领域知识定义了一些模板。通过实验将这些方法与一些优秀的问答模型进行了比较，实验结果证明了 BERT-BiLSTM-Att-CRF 模型有更优的性能。最后，在前述研究的基础上，设计并实现了一个基于知识图谱的问答系统，该系统支持领域知识图谱的导入，并使用 Neo4j 图数据库进行存储，同时在系统内实现了两种问答方法：基于语义模板的问答方法、基于 BERT-BiLSTM-Att-CRF 的问答方法。

本文实现的问答系统能够在一定程度上改善问答任务的准确率，更好地满足普通用户对于特定领域信息的查询需求。

关键词： 知识图谱；问答系统；深度学习；神经网络；自然语言处理

ABSTRACT

Knowledge Graphs(KG) can model and store domain knowledge well, but it still can't directly meet the need of common users. People are more accustomed to obtain answers through Question-and-Answer(QA) method. Therefore, There's an urgent need to establish Knowledge based question and answer system(KBQA).

Focused on the requirements above, this passage firstly combined the attention mechanism with the classical BERT-BiLSTM-CRF model for the reason that traditional methods based on BERT get low precision when answering Chinese questions. Therefore, we get the BERT-BiLSTM-Att-CRF model. Secondly, this passage also studies the QA method based on templates and defines some templates according to the knowledge from specific fields. This passage compares these methods with some outstanding models and the consequences prove that BERT-BiLSTM-Att-CRF has better performance. Lastly, based on the above studies, this passage designs and realizes a KBQA. The system permits the importation of KGs from specific fields and stores them by using Neo4j. In the system, two QA methods are working: QA based on templates, and QA based on BERT-BiLSTM-Att-CRF.

The KBQA realized by this passage can improve the precision of QA tasks to some extent, and can better meet the need of common users ti get information from specific fields.

KEY WORDS: Knowledge graph，Question and Answering System, Deep learning, Neural networks, Natural language processing(NLP)

目　　录

第一章　绪论

1.1　课题来源

从上世纪 60 年代起，早期的问答系统陆续出现，其中包括能够回答棒球比赛相关问题的系统以及能够回答月壤相关问题的系统。

从 90 年代起，问答系统的发展日趋成熟，出现了基于 Internet 的问答系统，以及能够回答使用多种语言提出的问题的问答系统。

当今社会，随着社会信息化的进程不断推进，在这个用户的信息获取需求不断提升的大背景下，传统的搜索引擎的劣势是：难以从输入的自然语言问句中分析出用户的真正意图；同时传统搜索引擎返回的信息往往需要进行二次筛选才能为用户所用。为了解决这些问题，同时更好的满足用户对于信息获取的需求，自动问答系统应运而生[1]。

问答系统需要从系统中带有的数据库来检索得到答案，这一来源通常是不同结构的文本，这也使得检索得到的答案中往往会包含较多的噪声数据。为了应对这一问题，需要完成大规模高质量的数据库构建。谷歌公司在 2012 年提出了“知识图谱”这一概念[2]，并通过知识工程等相关技术，使用异构信息完成知识库的构建，从而为问答系统提供了可靠的数据库。

近年来，随着大数据、自然语言处理等技术的快速发展，知识图谱在金融、旅游、购物等众多领域得到了广泛应用，这其中智能问答系统是知识图谱的主流应用场景之一[3]。

典型的 KBQA (Knowledge Base Question Answering，基于知识图谱的问答系统)[4]将自然语言问题转变为实体、属性、关系的三元组形式。利用转变得到的三元组来与用户输入问答系统的自然语言问句匹配，将匹配程度最高的三元组作为答案，以自然语言的形式返回给用户。

现有的基于知识图谱的问答系统相较于传统的搜索引擎在性能上有了长足的进步。随着深度学习技术的不断发展，以及这一技术在问答系统中的不断深入和广泛应用，KBQA 以及相关的技术都有了一定的进展。

当前，国内对于 KBQA 以及相关领域都有了一定的研究基础，但相关的研究仍然存在着一些缺陷，主要体现在下面两点：

第一，当前主流的成熟大规模知识图谱主要以英文数据作为载体，中文知识

图谱的发展相较而言仍然落后，而且存在着知识表示不完整、实体间关系错误等问题。

第二，中文与英文在语言特点上有着较大的差异。英文的表达方式相较而言更为单一，而且有着固定的格式可以遵循，此外，英文的单词之间天然带有空格作为分隔符，极大地减轻了问答系统的语义分析任务的难度；与英文相比，由于中文无论在语序还是同义词汇的使用上都更为灵活，而且作为语句组成基本元素的汉字之间也没有天然的分隔符，导致问答系统难以对词汇基于语义进行划分，这也为设计基于中文知识图谱的问答系统提出了更大的挑战[5]。

综上所述，基于中文知识图谱的问答系统及其相关技术仍然有着许多等待克服的问题。随着人工智能技术、知识图谱及其相关技术、自然语言处理技术等技术的发展，智能问答系统也正处于不断的发展之中。

基于中文知识图谱的问答系统具有重要的研究与应用价值，将在未来的科研与日常生活中得到更为广泛的应用。本研究旨在设计并实现一个支持领域知识图谱导入的 KBQA，并对现有的 KBQA 方法进行一定的改进。

1.2 国内外研究成果

目前关于 KBQA，国内外已有不少相关的研究成果，以下是该领域的研究成果按照时间进程的介绍：

问答系统雏形出现的时间可以追溯到 1950 年前后，当时提出的“图灵测试”可以归类为一项问答任务。

进入 60 年代，基于使用统计学语言问答方法的问答系统开始出现。此时的问答系统主要使用词类和词序。同时期，世界上第一个真正意义上的自动问答系统 ELIZA 问世[6]。

到了 70 年代，问答系统开始涵盖语义以及语言运用等领域，典型代表为 MARGIRE 系统、PAM 系统、SAM 系统等。

80 年代到 20 世纪末，问答系统从面向特定领域逐步发展为开放领域问答系统，这些问答系统的核心主要是需要人工预先制定的规则，这也导致了同时期的问答系统可扩展性较差。

进入 21 世纪，随着深度学习技术的出现与蓬勃发展，基于知识图谱以及基于网页链接搜索的问答系统相继问世，其中以 IBM 公司于 2011 年发布的 Watson 系统为典型，能够在较短时间内准确回答医学领域的问题。

相较于国外，国内早期在问答系统以及相关领域的研究成果相对而言较少，这一现像背后的原因，主要有以下两点：

一是国内在相关领域的研究起步较晚，缺乏足够的资料以及知识库等；

二是中文自身的语言特点。例如同一种语义可以由多种词汇或语句来表达，这对于支持中文问句的问答系统的设计提出了极大的挑战。

近年来，NLPCC IPPCOL KBQA 提供了一个大规模中文知识库，以其为基础，研究工作得以广泛开展[7]。

Shen 等[8]在提取自然语言问句关键词时，使用了 Bi-LSTM-CRF 模型，并在识别错误实体过程中加入编辑距离优化，用多粒度技术对自然语言问句与侯选属性完成编码，将 ABMGIM 模型融合进属性映射过程中，从而得到深层语义信息。

Dong 等[9]针对传统问答方法带有的实体错误识别问题，提出了一个从问句到知识图谱三元组的端到端解决思路。该方法首先使用 BM25 算法计算出候选的知识图谱三元组，然后利用 MFSMM 模型对得到的三元组序列进行排序，后续再计算其相似度，在这一过程中使用 Bi-LSTM 与 CNN 以实现语义级别和词级别的计算，最后按照计算出的相似度选出作为答案的知识图谱三元组。

当前主流的问答系统主要采用信息检索与语义解析的问答方法。关于信息检索及语义解析的问答方法介绍如下：

1.2.1 采用信息检索的问答方法

采用信息检索的问答方法主要从用户输入的自然语言问句以及问答系统自带的知识库之中提取特征进行匹配。

问答系统首先从用户输入的问答系统之中提取出命名实体，然后将提取到的实体与系统的知识库进行链接，得到候选答案集合。最后将候选答案与问句进行属性相似度比对，比对得分最高者作为最终答案返回给用户。

首先提出将信息检索的方法与 KBQA 结合起来的是 Bao 等人[10]。这一思路首先对用户输入的自然语言问句进行语法分析，从分析结果中获得核心词汇。之后，将用户输入的自然语言问句转换为与之对应的特征图，使用关键词检索系统的知识库，寻找与之匹配的三元组。最终将问句中提取到的特征以及检索到的三元组进行结合，得到答案。但是这种方法需要预先定义大量的人工规则，方法的可扩展性较差，难以在其他领域得到应用。

为了弥补上述不足，Jain 等人[11]使用 Embedding 方法将用户输入的自然语言

问句与检索到的候选答案映射为向量，计算两个向量的相似度，同时使用梯度下降方法对算法进行了优化，从而不再需要人工预先定义规则与构建特征。但是这一新的方法没有考虑多个语句序列之间的依赖关系。

Chen 等[12]在此基础之上，将神经网络应用于问答系统的设计与实现，所提出的 MCNN 模型利用卷积神经网络（Convolutional Neural network, CNN）编码自然语言问句与候选答案，并在模型之中考虑了句子信息，从而让系统能够对问句与候选答案之间的关联信息进行学习，最终检索到正确答案。

除此之外，由 Chang 等[13]提出的求解多约束问题的问答方法，这一方法使用新的数据集 ComplexQuestions。Jain 表的事实记忆网络模型，使用主语、谓语、宾语三元组来存储事实，并在多跳推理时使用记忆网络，从而极大地改善了性能。

Hsasn 等[14]提出的关系检测模型，在获取自然语言问句与知识图谱三元组的关联信息时，结合了注意力机制，能够将问句更准确地转化为向量，提高了模型处理复杂问题的能力。

1.2.2 采用语义解析的问答方法

采用语义解析的问答方法，首先需要将用户输入的自然语言问题转变为逻辑模板，再将逻辑模板转变为对应的查询语句，从而在知识图谱中检索答案。

传统的采用语义解析的问答方法需要人工预先定义的规则与模板，之后才能解析用户输入的自然语言问句，但这种方法需要过多的人工干预。

为了减少人工干预，使用查询图的语义解析方法应运而生。Yih 等[15]提出了 STAGG 模型，将语义解析过程转换为查询图的生成过程，最终答案由 lambda 算法排序得到。

在语义解析过程中，除了使用查询图，还有其他的选择。Xu 等[16]提出的 Graph-to-Sequence 模型，使用语法图对自然语言问句的信息进行表示，之后句法图使用 Graph2Seq 模型进行表示，上下文信息使用 Attention-RNN 获取，最终将自然语言问句转换为相应的逻辑表达式。

Hu 等[17]提出了基于状态转移的方法，通过对特定操作与状态转移模型进行预定义，复杂模型将被转化为语义查询图，之后利用 MCCNN 截取查询图的节点，从而不再需要人工预定义规则。

周博通等[18]提出的 InsunKBQA 系统，能够从多个层次上提取候选关系与自然语言问句中的语义信息，从而得到更为准确的答案。

Borders 等[19]提出的子图嵌入模型使用键值对存储知识图谱，在推理复杂问句时使用多个记忆模块，最终在知识图谱中检索答案。该模型可被用于特定领域推理，但在大规模通用知识图谱中的表现不尽如人意。

从上述材料不难看出：基于知识图谱的问答系统在中文与英文领域都有了长足的发展，其中中文自动问答系统的发展尤为迅速，不同领域的需求以及学者们的不断深入研究推动了中文问答系统的发展。

尽管如此，现阶段的中文领域问答系统仍然有着一定的不足之处，本研究将进一步对基于知识图谱的中文自动问答系统进行深入研究。

1.3 研究内容

基于知识图谱的问答系统根据自然语言问句，从系统知识库中选取与之匹配的三元组，作为答案返回给用户。

其工作流程可以做如下细化：用户向问答系统输入自然语言问句后，由问答系统从输入问句中抽取关系，之后从问答系统存储的知识图谱中检索与其相匹配的关系，最后利用从问句中提取出的实体与关系来决定最终答案，并返回给用户。

在上述步骤之中，实体抽取以及关系抽取是两个关键点。前者需要从自然语言问句中识别、提取出实体，并将提取到的实体链接到知识图谱之中，构建出与实体相关联的候选三元组序列；后者的任务是识别从自然语言问句中提取出的关系，具体说来，是从候选的三元组序列中筛选出匹配程度最高的一个。

本研究将问答系统分为实体抽取与向量映射两个部分：

实体抽取部分：使用 BERT-BiLSTM-Att-CRF 模型来实现对于命名实体的识别任务，利用能够较好地提取序列信息的 LSTM 对输入的自然语言问句编码，编码操作完成后，在应用注意力机制的基础上使用条件随机场（Conditional Random Field，CRF）对序列进行标注，对于标注后的问句，实体识别操作会更为便捷。对于识别结果模糊的实体。匹配任务使用编辑距离（Levenshtein Distance）完成。

向量映射部分：从语义层级以及词汇层级，对输入的自然语言问句和知识库中候选属性的信息进行提取。在提取出相应的语义信息后，利用 Concatenate 方法拼接不同层级的语义信息向量，得到新的问句信息向量以及候选属性集向量，二者间的相似度使用余弦距离进行计算，距离最小者最相似，将其作为问题答案。

本研究的具体工作可以概括如下：

（1）以 SPO（Subject，Predicate，Object）三元组为基础，构造一个用于问答系统的知识图谱数据库，该数据库能够支持领域知识图谱的导入。

（2）在向量映射部分，首先利用结合了注意力机制的 BiLSTM 模型从输入的自然语言问句和候选的属性中提取语义信息，再利用相似矩阵提取自然语言问句和侯选属性的词汇信息，将这两个层级的信息拼接后，计算其匹配程度，找出其中得分最高者。

实现采用上述步骤的中文 KBQA 系统，并对 KBQA 系统做测试，测试结果表明，本研究所提出的用于构建中文 KBQA 系统的 BERT-BiLSTM-CRF 模型是可行的，而且基于此模型所实现的 KBQA 系统的问答任务完成效果相较于采用传统方法的 KBQA 系统有所提升。

1.4 论文组织结构

本研究主要以基于中文知识图谱的问答系统作为研究对象。整篇论文共包含五个部分，具体内容如下所述：

第二章为实现基于中文知识图谱的问答系统需要用到的相关知识，以及实现该问答系统所需要的步骤，其中对于知识图谱及其相关技术、命名实体识别及其有关技术、属性映射及其有关技术做了着重介绍，并涉及了深度学习的相关理论。

第三章重点展示了本研究设计并实现的基于中文知识图谱的问答系统工作流程。第三章中，首先给出了对于系统的问题定义以及系统建模。整个问答系统可以分为两大部分：命名实体提取部分以及向量映射部分。之后介绍了命名实体提取任务所采用的经典算法及其不足之处、本文所使用的 BERT-BiLSTM-Att-CRF 模型的结构以及训练模型的过程，其中重点介绍命名实体提取任务中的链接环节。

第四章详细解释了本研究所使用的 BERT-BiLSTM-Att-CRF 模型以及用于对几个对照模型进行测试的评测指标，并给出了 BERT-BiLSTM-Att-CRF 模型与几个经典的模型在相同测试集下的运行效果，以及基于这几种模型所构建的中文 KBQA 系统在完成问答任务时的表现。其中对给出的模型的具体网络结构和工作流程给出了详细描述。`

第五章给出了对本研究设计并实现的基于中文知识图谱的问答系统的总结，探讨了本研究使用的 BERT-BiLSTM-Att-CRF 模型的优点以及不足之处，并在当前已经完成了的实验基础上进一步做出了未来研究方向的展望。

第二章　相关技术研究

2.1　基于知识图谱的问答系统综述

知识图谱（Knowledge Graph）这一概念在 2012 年由谷歌公司提出，属于有向图结构。传统搜索引擎由于存在匹配方式效率低下且难以理解用户输入问题的深层语义的问题，难以满足用户的需求，知识图谱正是为了解决这些传统搜索引擎存在的问题而提出的。

知识图谱实质上是由许多 SPO（Subject， Predicate，Object）三元组构成的。其基本形式主要可以分为两类，一类是**<实体、关系、实体>**类型的三元组；另一类是**<实体、属性、值>**类型的三元组。

知识图谱所存储的知识，按其来源主要可以分成两类：一类是半结构化的规范文本数据，主要来自百科文本；另一类是非结构化的自由文本数据，主要来自文件、评论、新闻等。

知识图谱的构建涉及到知识抽取、知识融合、知识加工等步骤[20]。

知识抽取：从来自不同数据源的原始数据中，抽取出能够构成知识图谱的三元组；

知识融合：将知识抽取过程中获得的异构信息整合，并消除歧义，保证知识库中知识的一致性；

知识加工：对知识库中已经存储的知识进行更新、推理等操作。

基于知识图谱的问答系统以系统中存储的知识图谱作为其答案来源，系统的核心任务是对自然语言问句进行语义分析和语义识别操作，这些输入问答系统的自然语言问句主要属于非结构化自由文本，但问答系统中的知识是以结构化存储的三元组，如何在这两者之间完成匹配是当前的该领域的一个难点。

在基于知识图谱的问答方法中，传统方法有采用语义解析的问答方法以及采用信息检索的问答方法。

随着深度学习方法及其相关技术的不断发展及其在问答系统中的不断深化应用，采用深度神经网络的问答系统已经能够获得令人满意的效果。

在本研究设计并实现的基于知识图谱的问答系统中，同时实现了基于语义解析的问答方法和采用深度学习的问答方法，并对传统的语义解析方法进行了一定的改进。

2.2 命名实体识别与 BiLSTM-CRF 模型

命名实体识别（Named Entity Recognition，NER）的任务，是对给定文本中涉及的命名实体进行识别。这里的命名实体包括姓名、地址等专有名词[21]。

NER 是自然语言处理领域的一个重要分支，也是实现基于知识图谱的问答系统中的关键环节，NER 任务的完成情况将直接影响到问答方法的准确度，并最终对问答系统的表现产生巨大影响。

传统 NER 方法由于其需要人工预先准备好的特征和知识，因此其可扩展性较差。随着机器学习及其相关技术的不断发展及其在问答系统领域的不断深化应用，NER 方法与机器学习技术也得以结合起来。

当前采用机器学习技术的 NER 方法有两个分支：一是将给定任务按照分类的方法进行实现；二是将给定任务按照序列化的方法进行标注。实验结果证实了后者具有更为良好的效果，这其中应用最为广泛的，是使用条件随机场(Conditional Random Field，CRF)[22]来完成序列化标注任务。

在采用了机器学习技术的问答系统之中，传统的方法由于需要人工预先提取特征并对相关知识进行预定义，其自动化程度相对有限。而且由于各个领域的知识互通程度较低，导致人工定义的特征与预定义知识难以在其他领域得到应用，需要人工重新提取特征并重定义知识。系统的可扩展性与可扩展性较差。

近年来，随着深度学习的不断发展及其在问答系统领域的广泛应用，原先需要人工预提取的特征可以使用深度神经网络实现自动化。在采用了深度学习技术的问答系统中，自然语言问句的特征由深度神经网络进行提取，再使用统计学方法完成序列标注任务。按照这一实现思路，Lample 等[23]提出了 BiLSTM-CRF 模型，由于 BiLSTM 能够对具有时序性的数据进行处理的特点，BiLSTM-CRF 模型使用 BiLSTM 提取自然语言问句的语义特征，序列标注任务由 CRF 完成。BiLSTM-CRF 模型由于其出色的表现，目前已经成为完成 NER 任务的主流模型。

长短期记忆网络(Long Short Term Memory network，LSTM)[24]是循环神经网络(Recurrent Neural Network， RNN)的一个衍生产物，引入了“门”的概念，从而能够处理 RNN 模型本身具有的梯度消失和梯度爆炸问题。上述的“门”共有三种，分别是输入门、输出门和遗忘门。

长短期网络具有记忆能力，可被用于对时序数据进行处理[25]。长短期网络拥有学习长距依赖关系的能力，但由于长短期网络结构中的激活函数以及反向传播部分的链式求导，长短期网络在使用时会带有梯度消失和梯度爆炸的情况，从而

影响其处理长距依赖关系的能力[26]。

BiLSTM-CRF 模型的工作过程，实质上是将 NER 任务转变为标注序列的任务。具体流程是把输入的自然语言文本划分为多个字符序列，之后再标注各个序列中的字符。标注过程采用主流的 BIO 标注法，使用“B”标注命名实体起始点；使用“I”标注命名实体剩余量；使用“O”标注非实体。例如，对于自然语言问句：“中国的全称是？”，对该问句进行标注后，得到的标注序列是：“B-Country，I-Country，O，O，O，O，O”，其中 Country 为实体类型，“中国”的实体类型对应为“Country”。

上述 BiLSTM-CRF 模型的结构示意图如下所示：

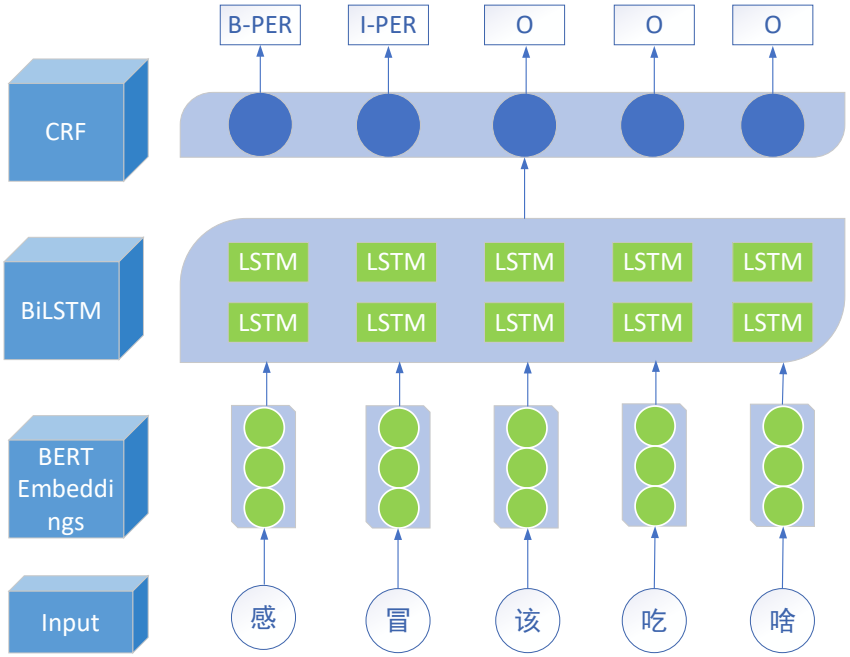

图 2-1 BiLSTM-CRF 模型结构示意图

其中(w_0, w_1, w_2)代表模型的输入序列。该输入可以是单词、字符等。(y_0, y_1, y_2)代表模型的输出序列，每个输出序列都是对应的输入序列的标签。序列输入到模型中后，首先经过 Embedding 层，在该层中输入序列被映射为相应的向量，之后向量将会被传输给神经网络进行进一步的运算。向量在输入神经网络后，被转换为矩阵形式，矩阵的大小与向量的序列长度和维度数正相关。之后，矩阵在进入 BiLSTM 模型后，经由前向计算与后向计算，分别得到该时刻的上文信息$\overleftarrow{h_t}$以及下文信息$\overrightarrow{h_t}$，对二者进行链接，即可得到一个完整的上下文信息$H_t = [\overleftarrow{h_t}, \overrightarrow{h_t}]$。

该矩阵随后在全连接网络中转换为一个标签矩阵。对该标签矩阵进行语义分析后，即可得到相应输入序列的输出标签。为解决序列之间的依赖关系问题，在模型中引入了 CRF 层，用于对语句层级的序列特征进行约束。其中 CRF 层使用

打分函数对输入的标签矩阵进行打分，该层使用的打分函数如下所示：

$$S(x,y) = \sum_{i=0}^{n} A_{y_i,y_{i+1}} + \sum_{i=1}^{n} P_{i,y_i} \quad (2\text{-}1)$$

等号右侧的 A 代表转换分数矩阵，该矩阵中的每个元素为标签转换后的得分，矩阵 A 在赋初值后，经过模型的迭代训练而最终达到最佳值。

等号右侧的 P 代表路径分数矩阵，该矩阵中的每个元素为单词与标签的匹配分数，矩阵 P 中的值由 BiLSTM 模型计算得出。

2.3 编码比较框架

属性映射过程作为实现一个 KBQA 的关键一步，其重点在于从问答系统的知识库中检索到与输入的自然语言问句匹配度最高的属性集合。

这其中文本相似度需要通过比较多个文本之间的相似度，并加以量化呈现[27]。在大量文本相似度计算方法中，字符串匹配是其中最简单的一种，该方法通过比较两个文本之间相同字符的个数，进而定义两个文本之间的相似度。该方法的不足之处在于：仅仅考虑了字符串层级的相似度，忽略了语义层级等关键信息。例如，使用字符串匹配将无法较好地计算“马铃薯”与“土豆”之间的相似度。

当前主流的计算文本相似度的方法是首先将文本进行编码，再对编码后的文本进行比较，按照这一思路构建的框架称为 Encoding-Compare 框架[28]。

其具体工作流程为：首先对两个待处理文本做语义解析，将二者降维编码为稠密向量，并映射到同一维度的向量空间，最后对两个向量的相似度利用计算相似度的算法求出。

本研究中问答系统的映射部分使用了 Encoding-Compare 框架，将输入系统的自然语言问句，连同问答系统知识库中的候选属性，采用编码器映射到相同向量空间中，计算出二者的相似度。

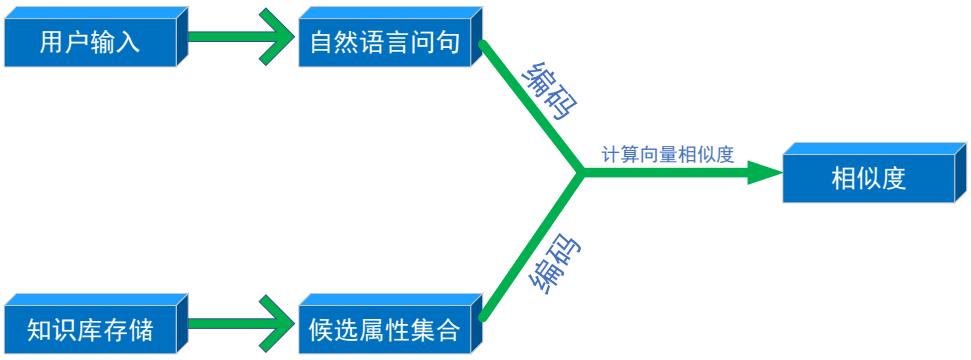

图 2-2 Encoding-Compare 框架工作流程图

2.4 计算文本相似度

在使用了深度学习方法后，两个文本间相似度的计算问题得以转化为两个向量间的相似度计算。

度量两个向量间相似度的常用方法有：曼哈顿距离、余弦距离、欧几里得距离等。上述方法从分别不同角度对向量间相似度进行考量，下面解释几个常用方法的计算过程：

对于两个向量 $U = [u_1, u_2, \dots, u_n]$ 和 $V = [v_1, v_2, \dots, v_n]$

（1）**欧几里得距离：**使用两向量在同一个向量空间的直线距离来反映向量的相似度，其具体计算过程如下：

$$distance = \sqrt{\sum_{i=1}^{n}(U_i - V_i)^2} \tag{2-1}$$

（2）**曼哈顿距离：**向量之间的距离使用两向量在坐标轴上的绝对距离来量化，具体的计算过程如下所示：

$$distance = \sum_{i=1}^{n}|U_i - V_i| \tag{2-2}$$

（3）**余弦距离：**向量间相似度使用两向量间的夹角计算得出，夹角越小则相应的向量对相似度越高；反之，夹角越大则相似度越小。

具体的计算过程如下所示：

$$distance = \frac{U \cdot V}{\|U\|\|V\|} = \frac{\sum_{i=1}^{n} U_i V_i}{\sqrt{\sum_{i=1}^{n} U^2}\sqrt{\sum_{i=1}^{n} V^2}} \tag{2-3}$$

上述三种量化相似度的方法中，不难发现：

欧几里得距离忽略了一对向量间不同维度的关联关系，因为转换为向量后的文本，其各个维度之间并非独立分布，它们之间存在着依赖关系。

曼哈顿距离将向量的不同维度之间的依赖关系纳入了考虑范畴，但当维度数较大时，其效果也难以令人满意。

余弦距离在计算过程中，将向量的不同维度之间的依赖关系纳入了考虑范

畴，同时避免了向量的长度对于最终输出的影响。此外，余弦距离是归一化的，因而其结果能够更为直观地反映两个向量的相似性。

2.5 注意力机制

深度学习模型在工作过程中，需要输入并处理大量数据，但是这些输入的数据中通常只有少部分数据对于输出的数据较为重要。注意力机制正是完成这一任务，即实现对于输入数据的重要性筛选的重要机制。根据输出向量的选择方式，注意力机制可以更进一步地细分为软注意力机制、硬注意力机制、键值对注意力机制、自注意力机制等。

软注意力机制使用注意力分布来对各个输入向量求加权和，从而实现对多个输入向量的融合；

硬注意力机制根据注意力分布来从多个输入向量中选择出其中的一个作为输出，所使用的选择方式主要有两种：选择得分最高的输入向量作为输出；根据随机采样的结果选择输出；

键值对注意力机制使用更一般化的键值对来代替软注意力机制和硬注意力机制中的输入向量。在键值对注意力机制中，使用查询向量和相应的键值来计算对应的注意力权重。在计算出输入数据的相应注意力分布后，使用该注意力分布和键值对中的键值完成加权融合计算。需要注意的是，若键值对的键值相同，则键值对注意力机制就退化为经典的注意力机制。

在自注意力机制中，使用的查询任务由输入的信息自动生成，而非选择一个与具体任务相关的查询向量。也即，模型在获取到输入信息后，自动地根据输入信息自身来决定其中相对重要的部分。自注意力机制往往使用查询-键-值的形式，在获取到输入信息在不同空间的表示后，即可得到一个注意力输出向量。

第三章　基于 BERT-BiLSTM-Att-CRF 的问答模型

3.1　系统定义与构建

KBQA 需要实现的功能是利用系统的知识图谱数据库中的知识三元组，对用户输入的自然语言问句进行回答。其中的关键部分在于分析用户输入的自然语言问句的语义，以及按照分析出的语义从知识图谱数据库中检索到匹配的答案。

本研究所采用的基于语义分析的 KBQA 可以分为两个主要部分，分别是实体提取部分与向量映射部分。

实体提取部分：作为整个 KBQA 中的重要环节，其任务是对用户输入的自然语言问句进行命名实体识别，提取出语句中的命名实体，并将提取到的实体与系统中存储的知识图谱进行链接。通过链接步骤在知识图谱库中筛选出与用户输入的自然语言问句有关的候选知识三元组。实体提取部分可以进一步细化为两个子部分，其一是命名实体识别部分，其二是命名实体链接部分。

向量映射部分：对上述筛选出的候选知识三元组进行排序，挑选出关联程度最高的知识三元组，将其作为答案返回。

整个 KBQA 的工作流程可以概括如下：

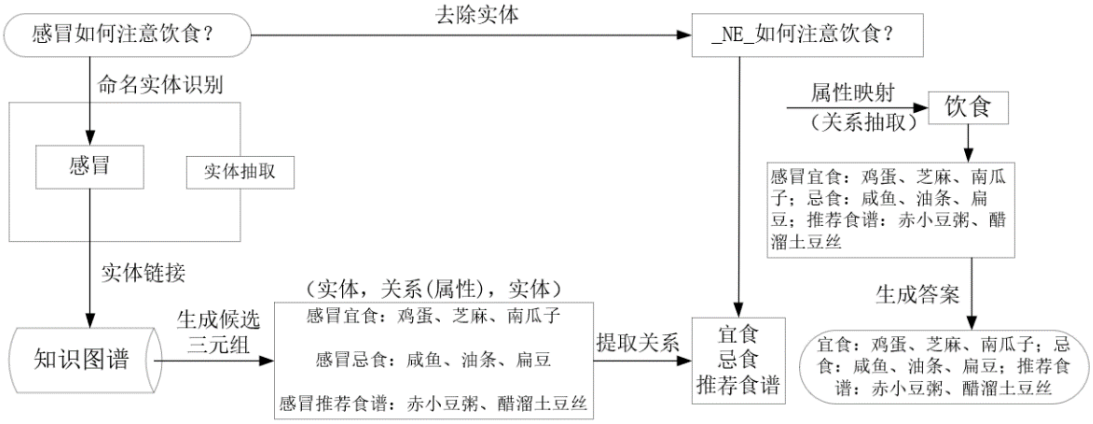

图 3-1 KBQA 工作流程图

本文为系统做出如下定义：给定一个自然语言问句

$Q = (w_0, w_1, w_2, \dots, w_p, \dots, w_q, \dots, w_n)$，等号右侧的每个$w_i$表示自然语言问句之中的一个字符，下标$n$代表自然语言问句的长度。

定义从自然语言问句Q中提取出的命名实体为$Entity = \{w_p, \dots, w_q\}$。之后将提取到的实体与问答系统存储的知识图谱进行链接，得到与该实体有依赖关系

的知识三元组。

定义从知识图谱数据库中筛选出的侯选属性集合为$R = \{r_1, r_2, \dots, r_n\}$，等号右侧的每个$r_i$表示一个属性，下标 n 代表属性集合的大小。向量映射部分的关键任务是从侯选属性集合 R 里筛选出与输入的自然语言问句 Q 相似度最高的属性，具体的筛选过程如下所示：

$$R^+ = \arg max S(Q; R) \tag{3-1}$$

上式中等号左侧的R^+表示筛选出的属性，等号右侧的$S(Q;R)$表示分数计算函数，其含义是当输入的自然语言问句为Q，侯选属性集合为R时对应的分数。具体的比较过程使用余弦距离来完成。

3.2 命名实体识别

经典的 BiLSTM-CRF 模型兼具深度学习问答方法与传统机器学习问答方法的优点，在提取英文问句中的命名实体时有着出色的表现。该模型的 BiLSTM 部分用于自动对输入的自然语言问句进行语义信息提取，CRF 部分用于对 BiLSTM 部分的输出结果进行约束，从而使输出的序列标注更为理想。但这一模型在被用于提取中文问句中的命名实体时，其结果并不能够令人满意。

经典的 BiLSTM-CRF 模型在编码自然语言问句时，仅仅将文本时序信息纳入考虑范畴，忽略了文本的局部信息以及单词之间的依赖关系。

除此之外，受中文本身的语言特点影响，中文对于同一语义有着多种不同的表达方式，而且句法灵活性相较于英文更高，结构更为复杂。且英文由于自身的语言特点，其句子中本身即带有分割词汇的空格，部分特殊名词的首字母会作大写处理，这些特点都极大的简化了命名实体识别任务。

中文由于汉字之间缺少相应的分隔符号，在加大了实体识别任务的难度的同时，可能会有歧义划分的问题出现，例如对于“苹果”一词，既可以将其解读为水果的一种，也可以将其解读为苹果公司，从而产生了歧义。

3.3 命名实体识别任务的数据预处理

在本研究所设计实现的 KBQA 系统之中，每个由用户输入的自然语言问句中都带有一个实体，这个实体也即是命名实体，其在问答系统的知识库中也存在。

在命名实体识别任务中，只需要对用户输入的自然语言问句中的单词进行实体判别，无需划定每个单词所属的实体类别。

本研究中使用 BIO 标注法来部分标注 NER 任务的数据，从而简化模型的训练过程。该标注法使用“B”标注实体起始部分，使用“I”标注实体剩余部分，使用“O”标注非实体部分。

3.4 模型改进与实现

作为端到端的深度学习模型，经典的 BERT-BiSLSTM-CRF 模型将人工提取特征的过程自动化完成。通过模型中 Bert 模块的 transformer 强大的特征提取功能，在经过预训练后，模型能够获得与单词相对应的嵌入式表示。相较于 BiLSTM+CRF 模型，结合了 BERT 后，模型的特征提取能力有了 极大的改善。本研究在经典的 BERT-BiLSTM-CRF 模型的基础上，进一步融入了注意力机制，在原有模型的 BiLSTM 层后加入了一层 Attention 层，用于在模型中使用注意力机制，从而根据不同的上下文信息的重要性分配不同的权重，进一步获取输入的自然语言问句的潜在特征。改进后的模型即为本研究所使用的 BERT-BiLSTM-Att-CRF 模型，如下图所示。

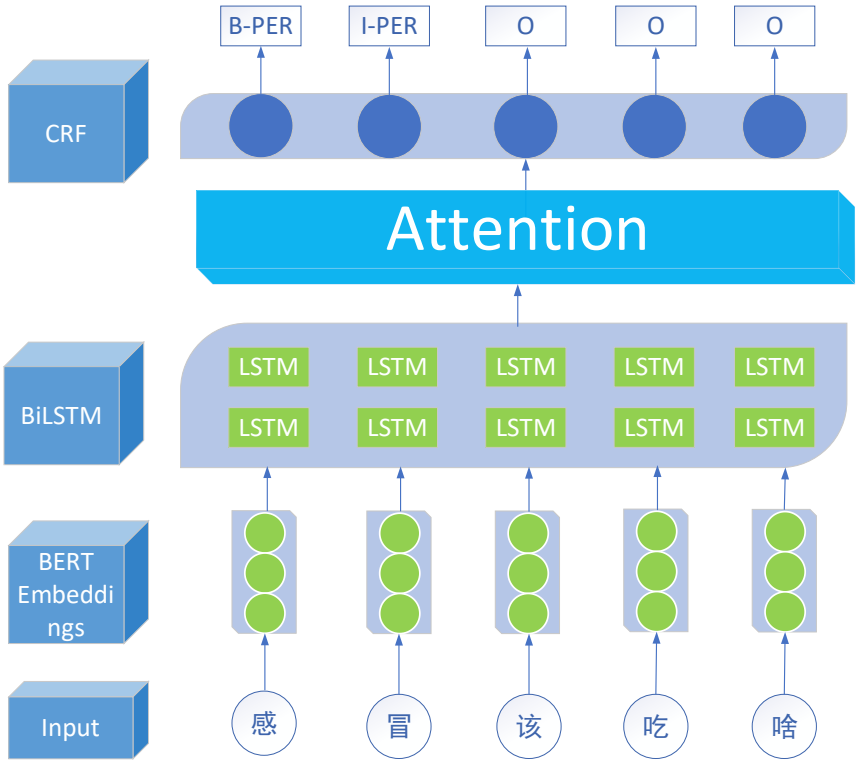

图 3-2 BERT-BiLSTM-Att-CRF 模型结构示意图

由于经典的 BERT-BiLSTM-CRF 模型在回答中文自然语言问句时，其表现相对而言较差。针对模型在这一方面的不足之处，在分析了模型内部的结构特

点后，为了优化模型的相关表现，主要进行了以下三个方面的改进：

（1）将经典模型中的实体链接任务转变为字符串相似度匹配任务。

为了改善 KBQA 任务中实体链接环节的任务完成情况，本研究将相应的实体链接任务转化为字符串之间的相似度匹配任务。如果上游的 NER 模型抽取出的实体可以和知识图谱中的实体成功链接时，则不对该实体进行任何处理；而当 NER 抽取出的实体无法和知识图谱中的实体链接时，则计算出抽取得到的实体和知识图谱中的相似文本之间的距离，根据距离的大小筛选出知识图谱中的对应实体。

（2）将注意力机制与经典的 BILSTM 网络融合起来。

在语义层次上，本研究在双向长短期记忆网络 BiLSTM 的基础上，融入了注意力机制。结合了注意力机制的模型将在语句层级上分别对输入的自然语言问句和候选属性进行语义解析，并将语义解析得到的结果以向量形式映射到相同的向量空间中。在使用向量形式对问句和候选属性进行表示前，首先需要从问句中删去存在的命名实体，通过这一步骤，问句中的主题实体对于属性映射的影响会有所降低，而强模型的计算量以及计算的复杂度都会相应的有所下降。

（3）使用字符层级的语义表示。

为了解决可能出现的未登录词（out-of-vocabulary，OOV）问题，本文使用了字符层级表示，并将该表示应用于语义层次。使得 OOV 问题得以解决，同时序列信息的获取也实现了最大化。

通过上述三个方面对经典的 BERT-BiLSTM-CRF 模型进行改进，改进后的模型在回答中文自然语言问句时，其表现相较而言有了一定的改善，效果如下：

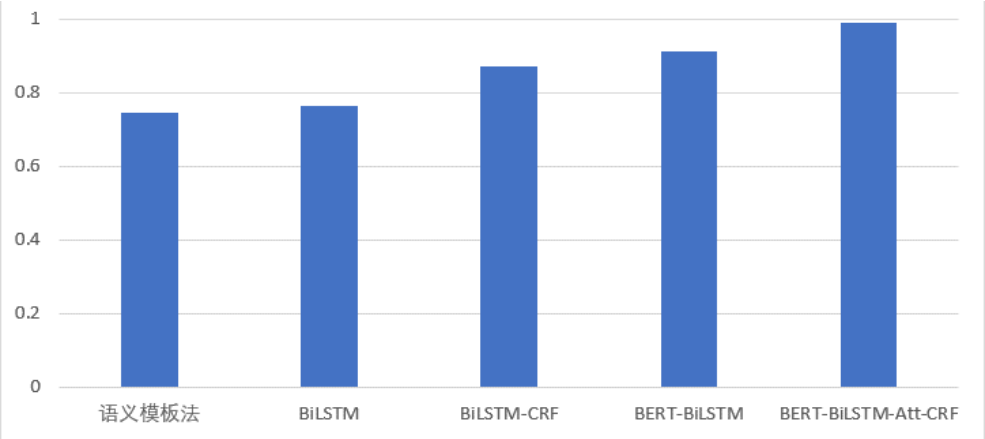

图 3-2 各模型准确率对比

为了进一步提高使用中文的 KBQA 完成命名实体识别任务的效果，本研究在经典的 BiLSTM-CRF 模型[30]的基础上，加入了双向编码器表示(Bidirectional Encoder Representation from Transformers，BERT)部分。整个模型

融合了 BiLSTM、BERT、注意力机制和 CRF 各自的优点，从而能够高效地实现命名实体识别[31]。

本研究使用的基于知识图谱的问答系统，首先预处理用户输入的自然语言问句，之后利用 BERT-BiLSTM+Att+CRF 模型来提取预处理后的自然语言问句中的实体，然后由系统匹配并分析已有的查询模板与输入的问句，按照改进的朴素贝叶斯算法来对问题做分类工作。问句被分类后，可以被用于查找匹配的查询模板。提取到的实体与实体关系将被嵌入到 Cypher 语句的结构中。

通过上述步骤，查询语句就可以被用于从问答系统存储的知识图谱中检索到答案，并将其返回给用户。

本研究设计并实现的基于知识图谱的问答系统分别实现了基于语义模板的问答方法以及改进的基于 BERT-BiLSTM-Att-CRF 的问答方法。

3.4.1 基于深度学习的问答方法

由于实际上的系统设计普遍含有模块化和结构化的特征，本系统采用的 BERT-BiLSTM-Att-CRF 模型自下而上分别由四个层次组成，分别是 BERT 层、双向 LSTM(Bidirectional LSTM，BiLSTM)[33]层，注意力层以及 CRF 层。其中，BERT 层用于充分利用上下文信息，并直接对原始数据进行处理，BiLSTM 通过调整输入门、输出门和遗忘门来从文本中提取深度语义实体。注意力层按照不同的上下文信息的重要性来分配不同的权重，从而进一步获取输入的自然语言问句的潜在特征。CRF 层则用于对注意力层权重分配后的结果进行解码操作，并对最终的表示序列进行优化。以 BERT-BiLSTM-Att-CRF 模型为基础，本研究所设计并实现的基于知识图谱的问答系统得以实现。

（1）预处理模块

对输入的自然语言文本进行表示是 BERT-BiLSTM-Att-CRF 模型的核心功能。文本中的每个词汇经由多层嵌入与一系列数字变换而得到表示。经过上述表示过程，将得到的每个词汇的最终语义表示结果输出。

模型的关键之处是使用注意力机制[34]来对文本进行处理。其基本原则是计算词汇之间以及词汇在整个句子之中的依赖关系与重要性。该过程用表达式表示如下：

$$Attention(K,Q,V) = softmax\left(\frac{K^T Q}{\sqrt{d_k}}\right)V \quad (3\text{-}2)$$

其中：K, Q, V是输入的词汇向量，d_k是输入的词汇向量的维数。

模块的输入由添加的相应的词汇向量表示得到，其中包括符号嵌入、分词嵌入以及位置嵌入[35]。同其他的语言模型相比，经过预处理的 BERT 模型能够充分利用词汇的上下文信息，同时获取到词汇的最佳分布式表示[36]。

（2）隐藏层

通过在经典的 LSTM 模型中引进能够识别长距离与短距离信息的记忆单元和限制门，从而使其能够解决梯度消失问题。

该模型在某些方面仍然有一些有待弥补的差距。为了提升模块的效率，可以对模型中的限制门做改进。

隐藏层的输出的具体计算步骤如下所示：

$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i) \tag{3-3}$$

$$f_t = \sigma\left(W_f[h_{t-1}, x_t] + b_f\right) \tag{3-4}$$

$$O_t = \sigma(W_o[h_{t-1}, x_t] + b_o) \tag{3-5}$$

$$\tilde{c}_t = i_t \times \tilde{c}_t + f_i \times c_{t-1} \tag{3-6}$$

$$h_t = O_t \times \tanh(c_t) \tag{3-7}$$

其中：W为相邻两层之间的联系，其中W_i、W_f、W_o都表示从输入层到隐藏层的权重矩阵；

b代表偏移向量，其中b_i、b_f、b_o都表示隐藏层的输入门的偏移向量；

c_t为记忆单元，$\tilde{c}_t$代表 t 时刻时记忆单元的状态；

$tanh$是两个不同的神经元激活函数；

i_t，f_t，O_t分别代表模型中的输入门、遗忘门以及输出门；

h_t代表 t 时刻时隐藏层的状态，也代表了此时的输出态。

考虑到经典的 LSTM 模型仅能获取到当前词汇的上文信息，而 NER 任务需要获得当前词汇的上下文信息。为了能同时获得上下文信息，本研究采用 BiLSTM 替代原有的 LSTM 完成文本编码任务。在 BiLSTM 结构中，前向 LSTM 用于获得 t 时刻词汇的上文信息$\overleftarrow{h_t}$，后向 LSTM 用于获得 t 时刻词汇的下文信息$\overrightarrow{h_t}$，二者链接后得到一个完整的上下文信息$H_t = \left[\overleftarrow{h_t}, \overrightarrow{h_t}\right]$。

（3）标签生成模块

对于输入序列$x = (x_1, x_2, \dots, x_n)$，假定 P 是从 Bi-LSTM 网络的输出中得到的分数矩阵，且对于预测序列$y = (y_1, y_2, \dots, y_n)$，假定最终得分可以定义为：

$$S(x, y) = \sum_{i=0}^{n} A_{y_i, y_{i+1}} + \sum_{i=1}^{n} P_{i, y_i} \tag{3-8}$$

其中：A 代表从标签 i 到标签 j 的过渡矩阵。

定义在所有可能的标签序列中，softmax 生成序列 y 的可能性如下：

$$P(y \mid x) = \frac{(e^{s(x,\ y)})}{(\sum_{\tilde{y} \in Y_x} e^{s(x,\ \tilde{y})})} \tag{3-9}$$

其中: Y_x代表输入序列x的所有可能的标签序列。

基于最大对数概念，系统可以调整网络参数，输出更有效的标签序列。在解码过程中预测输出序列，将分数最高者作为预测输出序列，用作 CRF 的输出。

根据命名实体识别规则，词汇被划分为四个类别：

B-PER，I-PER，E-PER，O。

其中 B-PER 表示字符位于实体字符界限的起始处；I-PER 表示字符位于界限中部；E-PER 表示字符位于界限末尾；O 表示字符与实体无关。

（4）关系提取模块

关系提取模块是问答系统中帮助问答系统理解查询语句的一个重要部分。

本研究中，使用朴素贝叶斯分类算法实现简单问题与模板间的匹配。朴素贝叶斯分类算法基于条件概率、贝叶斯定理和独立性假说。其详细处理过程如下：

$$P(c_i \mid x,\ y) = \frac{P(x,\ y \mid c_i) P(c_i)}{P(x,\ y)} \tag{3-10}$$

其中：x和y代表特征变量，c_i代表分类标签，$P(c_i \mid x,\ y)$代表特征变量为x和y时，被分类到c_i类的概率。

（5）Cypher 语句生成模块

Cypher 语句生成模块的主要功能是基于查询实体 E 和实体间关系 R 来构造 Cypher 查询语句，利用得到的查询语句检索答案[37]。以如下的查询语句为例：

```
match(n : e) - [re : r] → (s)
return  s
```

上述语句中：n : e 代表使用实体名 e 赋值实体 n；

语句 re : r 代表事业关系名 r 赋值关系联系 re[38]。

由上述语句生成查询语句，利用关系 r 与实体 e 来查询系统的知识图谱库，并返回符合限制的三元组[39]。

3.4.2 基于语义模板的问答方法

为了更好地展示 BERT-BiLSTM-Att-CRF 模型在完成问答任务时的表现，本研究在实现的问答系统中同时实现了基于语义模板的问答方法，并在问答系统中构建了相应的模板库。在用于对照的语义模板法中，关系匹配与问句分类任务由改进的朴素贝叶斯算法完成，算法的输出结果用于查询模板匹配，其中提取到的实体与关系被转换成 Cypher 语句。

用于KBQA的语义模板法使用预先定义好的模板来和用户输入的自然语言问句进行匹配，从而得到相应的查询语句。语义模板法首先需要构造出与知识库相对应的问题模板，以及与问题模板相对应的查询模板。模板构建完成后，对于用户输入的自然语言问句，首先将其与模板库中的问题模板进行匹配，得到问题模板后进一步去查找对应的查询模板。在这之后，将查询模板实例化，也即从自然语言问句中抽取出相应的语义词汇，对模板进行填充，从而实现最终的查询[32]。

3.5 实验环境与实验数据集介绍

上一节介绍了实现一个中文 KBQA 系统所需要完成的两个重要环节：命名实体提取及向量映射。

为了验证本研究所设计并实现的中文 KBQA 的问答效果，本章使用中文医药知识图谱问答数据进行实验验证，并针对问答结果给出相应的分析。

软硬件项目	说明
CPU	Intel(R) Core(TM) i7-8550U CPU @ 1.80GHz 2.00Ghz
内存	16G
数据库	Neo4j
操作系统	Windows10
开发语言及环境	Python，Pytorch，Pycharm

表 3-1 硬件与软件环境

本文所使用的数据集主要来自垂直性医药网站，从这些医药网站所获得的数据包含 7 个实体类别，共计 44111 个实体；10 个关系类别，共计 294149 个关系。数据集中包含的实体与关系的中文含义、数量以及实例如下表所示：

实体类型(7 项)	中文含义	数量	示例
Check	诊断检查项目	3,353	心电图
Department	医疗科目	54	皮肤科
Disease	疾病	8,807	感冒
Drug	药品	3,828	维 C 银翘胶囊
Food	食物	4,870	赤小豆粥
Recommend_Drug	推荐药品	17,201	阿司匹林肠溶胶囊
Symptom	疾病症状	5,998	头晕
Total	总数	44,111	

表 3-2 实体类型分类

关系类型(10 项)	中文含义	数量	举例
acompany_with	疾病并发症	12,029	<过敏性休克,并发症,水肿>
belongs_to	属于	8,844	<妇科,属于,妇产科>
common_drug	疾病常用药品	14,649	<过敏性肺炎,常用,阿奇霉素片>
do_eat	疾病宜吃食物	22,238	<胸椎骨折,宜吃,黑鱼>
drugs_of	当前在售药品	17,315	<青霉素 V 钾片,在售,通药制药青霉素 V 钾片>
has_symptom	疾病症状	5,998	<早期乳腺癌,疾病症状,乳腺组织肥厚>
need_check	疾病所需检查	39,422	<感冒,所需检查,血常规>
no_eat	疾病忌吃食物	22,247	<唇病,忌吃,杏仁>
recommand_drug	疾病推荐药品	59,467	<感冒,推荐用药,消炎片>
recommand_eat	疾病推荐食谱	40,221	<感冒,推荐食谱, 葱蒜粥>
Total	总数	294,149	

表 3-3 关系类型分类

由于本研究所使用的数据来自数个医疗信息网站，从众多半结构化或非结构化文本中筛选出来，其中所获取的数据不可避免的带有噪声部分。

对于获取到的数据中的噪声，采用降噪方法去噪声后最终得到了本研究使用的数据集。

本研究使用字符与语句联合注释的命名实体识别方法，采用 BIO 标注法对实验数据集进行标注。标注步骤的任务是对用户输入的自然语言问句中的命名实体进行识别。本研究采用三元组反向标注的方法来构造命名实体识别任务的数据集，将三元组中的首位元素确定为自然语言问句的命名实体，这些实体一律采用 PER 表示，非实体部分采用 O 表示。如下是几个问句标注实例：

自然语言问句	标注结果
感冒不宜吃什么？	B-PER B-PER O O O O O O
肝病吃什么药？	B-PER B-PER O O O O O
耳鸣应该吃什么？	B-PER B-PER O O O O O O

表 3-4 问句标注样例

3.6 实验评价指标

本文使用精确率(Precision)、召回率(Recall)\准确率(Accuracy)和 F1，共 4 个指标来对 KBQA 的结果进行评价。

其中精确率代表问答系统能找到答案的任务中，输出答案正确的样本比例；

召回率代表系统正确回答的问答任务比例；

准确率代表系统成功回答的任务比例；

F1 用于平衡精确率与召回率两个指标。

四个指标的详细计算过程如下所示：

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (4\text{-}1)$$

$$Precision = \frac{TP}{TP+FP} \quad (4\text{-}2)$$

$$Recall = \frac{TP}{TP+FN} \quad (4\text{-}3)$$

$$F1 = \frac{2\times Precision\times Recall}{Precision+Recall} \quad (4\text{-}4)$$

在实验开始前，首先对模型中的参数进行设置，以便使模型的训练效果达到最佳，具体的参数设置如下所示：

参数名称	值
Batch-size	128
Dropout-rate	0.20
Embedding-dim	300
Epoch	100
Hidden-dim	400
Learning Rate	0.005
Kernel-size	(3,5,7)

表 3-5 模型参数设定

3.7 问答系统性能测试分析

使用上一节展示的模型参数设定对本研究使用的 BERT-BiLSTM-Att-CRF 模型，以及用于对照试验的 BiLSTM 模型、BiLSTM-CRF 模型、BERT-BiLSTM 模型和语义模板法进行训练。

训练过程中，在设定了不同的迭代次数后，记录其训练后的 F1 值的变化，将得到的训练过程中的 F1 值变化数据汇总成表，选择其中迭代 0 次至迭代 50 次训练得到的 F1 值，汇总成如下表：

采用模型	0	10	20	30	40
BERT-BiLSTM-Att-CRF	0	0.7431	0.8726	0.9231	0.9162
BERT-BiLSTM	0	0.5113	0.7516	0.7824	0.8217
BiLSTM-CRF	0	0.4212	0.6713	0.6993	0.7233
BiLSTM	0	0.4333	0.6495	0.6872	0.6941
语义模板法	0	0.3952	0.5776	0.601	0.6391

表 3-6 各模型迭代次数与对应 F1 值

选择迭代训练 0 次至迭代训练 100 次得到的 F1 值，绘制得到的折线图如下：

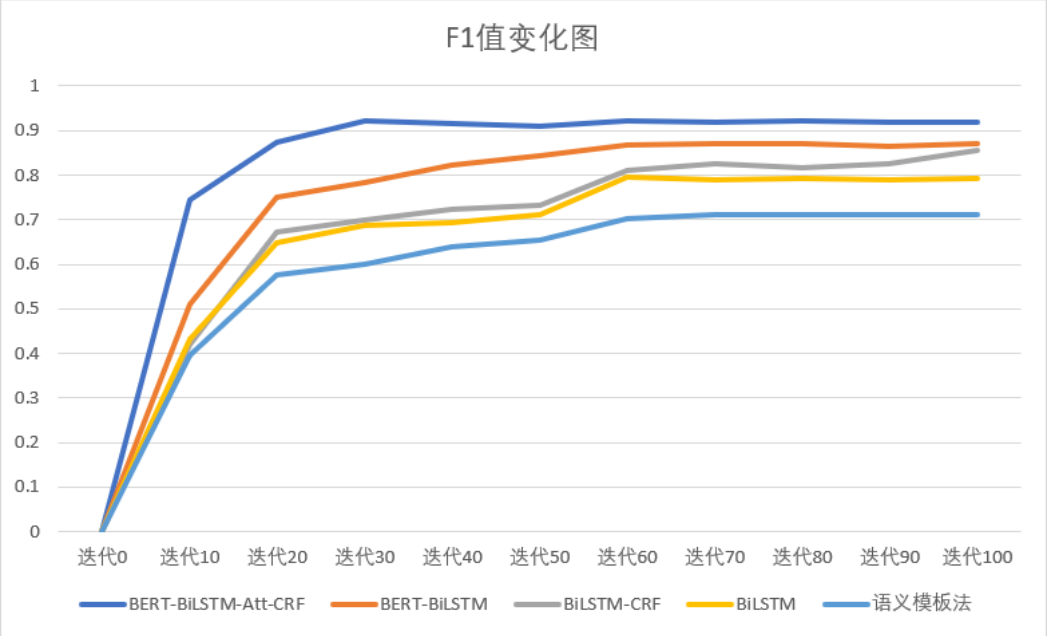

图 3-3 各模型训练过程中 F1 值变化

在对上述四个模型完成训练任务后，对这几个模型进行测试，测试任务使用相同的测试数据集，对各个模型的问答任务解决能力进行测试，任务过程中记录各个模型的回答成功次数、回答无效次数、答案检索失败次数进行统计，计算出每个模型在完成问答任务时的准确率与 F1 值，将评测指标汇总成图表，得到的各个模型的准确率与 F1 值变化如下：

采用模型	准确率	F1 值
语义模板法	0.744	0.698
BiLSTM	0.762	0.754
BiLSTM-CRF	0.869	0.857
BERT-BiLSTM	0.913	0.921
BERT-BiLSTM-Att-CRF	0.99	0.954

表 3-7 各个模型准确率与 F1 值

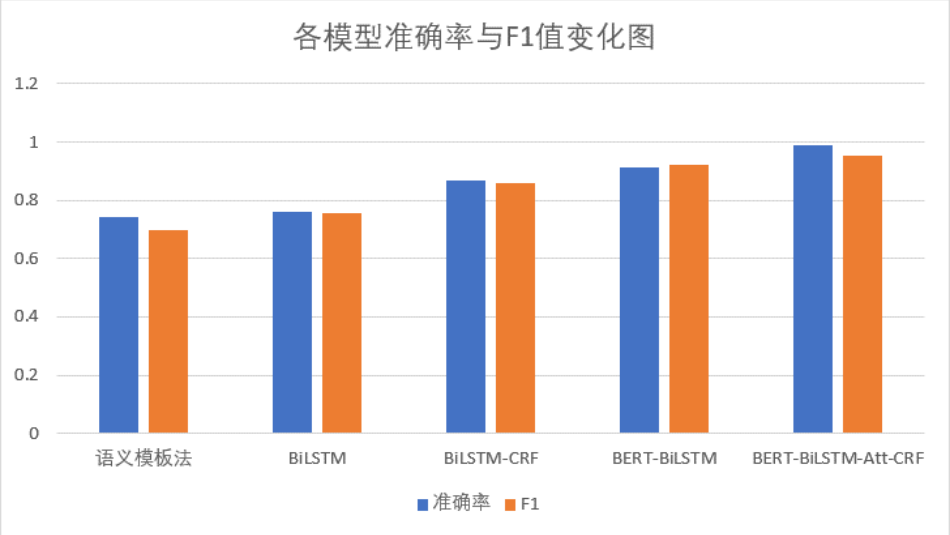

图 3-4 各模型准确率与 F1 值变化图

从上述实验中得到的数据，不难看出无论是采用准确率作为评测指标，还是采用 F1 值作为模型表现的评测指标，相较于其他经典的模型，本文所使用的 BERT-BiLSTM-Att-CRF 模型都达到了令人满意的效果。

第四章　基于知识图谱的问答系统的实现

4.1　系统设计

根据第三章提出的 BERT-BiLSTM-Att-CRF 模型，以及用于对照实验的基于语义模板的问答方法，本文设计并实现了一个基于知识图谱的问答系统，具体的设计过程叙述如下：

（1）知识库的设计：

本文设计的系统使用领域知识图谱作为搜索依据，领域知识图谱的数据使用 json、csv 等文件格式进行存储，并利用 Neo4j 图形数据库对 json、csv 等文件进行转换与处理，得到相应的图形数据库，即本系统使用的知识库。

图数据库使用图结构来对数据进行存储与处理，且其关系遍历操作的执行效率要远远高于关系数据库使用的表连接算法。另一方面，图数据库的面向图结构的分析和模式匹配、采用的 Cypher 查询语言，要比关系数据库更为直观、简洁。

（2）问答方法设计：

本文设计的 KBQA 分别使用了基于语义模板的问答方法和基于 BERT-BiLST

M-Att-CRF 的问答方法，两种算法的设计过程叙述如下：

基于语义模板的问答方法：

该方法使用预定义的模板来对用户输入的自然语言问句进行匹配，并根据匹配结果生成对应的形式化查询语句。

该方法的具体工作流程可以分为：命名实体识别、语义模板匹配、关系匹配、答案类型匹配、排序几个步骤。其中，模板匹配为该问答方法的核心部分，如下图所示：

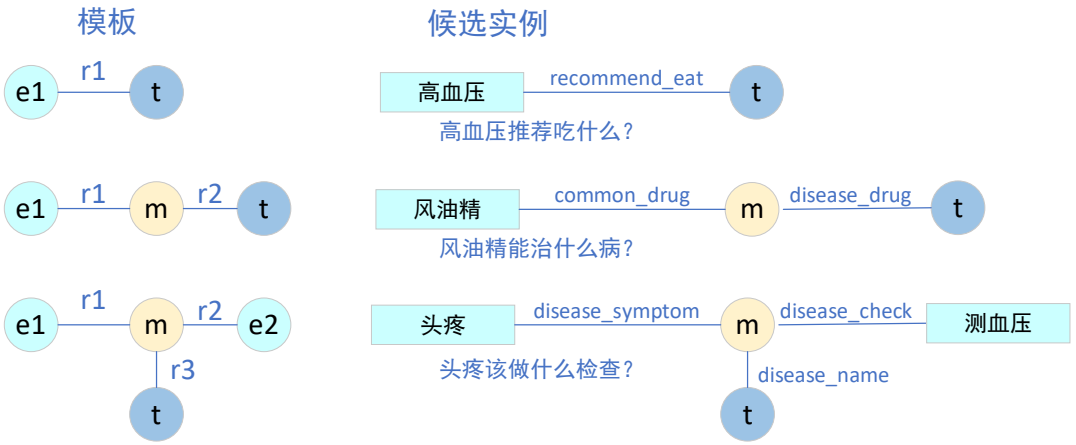

图 4-1 各模型准确率与 F1 值变化图

上图中，e 表示实体，m 表示中间变量，r 表示关系，t 表示答案。

用户输入的自然语言问句经过实体识别得到其中的实体，根据涉及到的实体使用图中的#1、#2 或#3 形式的模板进行匹配，每个模板将从系统的知识库得到一组查询结果。后续的关系匹配过程将从问句中识别出的关系，与从知识库查询出的关系进行匹配，不匹配的关系将被删除，最终剩下的查询结果为候选实例。

基于 BERT-BiLSTM-Att-CRF 的问答方法：

该方法基于上一章提出的 BERT-BiLSTM-Att-CRF 模型，模型首先使用 BERT 模型获取用户输入的自然语言问句中的实体向量，并提取文本的特征；之后由 BiLSTM 学习上下文的特征信息并做命名实体识别；在 BiLSTM 模块后，使用注意力机制对得到的结果分配权重并进行筛选；最后由 CRF 层对上一环节的输出序列进行处理，结合 CRF 自身存储的状态转移矩阵和相邻之间的标签，最终能够得到一个最优的结果序列。

（3）对于中文自然语言问句的处理：

主流的问答方法，如基于 BERT-BiLSTM-CRF 模型的问答方法等，在用于中文问题回答时，其效果相对而言较差。

本文使用的 BERT-BiLSTM-Att-CRF 模型在结合了注意力机制后，能够按照迭代训练得到的注意力权重分配，赋予问句中的重要实体与关系更高的权重，从

而重点关注问句中涉及到的命名实体和关系，更好地对用户输入的中文问句进行实体识别和特征提取。

4.2 系统实现

在系统实现过程中，涉及到了：数据处理、算法实现与改进、界面设计。

（1）数据处理：

对于本文设计的问答系统所使用的数据文件，如 json 文件，首先需要将其存储到图数据库中。这一过程通过使用 Neo4j 图数据库提供的函数，按照导入文件的路径读取文件内容，并按照其内容建立对应的知识图谱节点与关系。

```
'''建立知识图谱节点'''
def create_node(self, label, nodes):
    count = 0
    for node_name in nodes:
        node = Node(label, name=node_name)
        self.g.create(node)
        count += 1
        print(count, len(nodes))
    return
```

图 4-2 知识图谱节点的建立

（2）问答方法实现：

基于语义模板的问答方法：

该问答方法的核心部分在于语义模板的构建，本研究采用人工预先定义模板的方式，所定义的模板形式如下图所示：

```
self.accompany_que_words = ['并发症', '并发', '一起发生', '一并发生', '一起出现',
                            '一并出现', '一同发生', '一同出现',
                            '伴随发生', '伴随', '共现']
self.belong_que_words = ['属于什么科', '属于', '什么科', '科室']
self.cause_que_words = ['原因', '成因', '为什么', '怎么会', '怎样才', '咋样才',
                        '怎样会', '如何会', '为啥', '为何', '如何才会',
                        '怎么才会', '会导致', '会造成']
self.check_que_words = ['检查', '检查项目', '查出', '检查', '测出', '试出']
self.cure_prob_que_words = ['多大概率能治好', '多大几率能治好', '治好希望大么', '几率',
                            '几成', '比例', '可能性', '能治', '可治', '可以治', '可以医']
self.cure_que_words = ['治疗什么', '治啥', '治疗啥', '医治啥', '治愈啥', '主治啥',
                       '主治什么', '有什么用', '有何用', '用处',
```

图 4-3 语义模板的定义

模板定义完成后，用户输入的自然语言问句首先将用户输入的中文问句和模板库中的问题模板进行匹配，根据匹配的问题模板生成相应的查询模板。根据查询模板，按照从问句中抽取出的语义文本对查询模板进行填充，得到最终的完整查询语句。使用该语句即可从知识库中得到候选答案，对候选答案进行打分排序得到最终答案并返回给用户。

基于 BERT-BiLSTM-Att-CRF 的问答方法：

该问答方法在经典的 BERT-BiLSTM-CRF 模型中加入了注意力机制，从而能够对 BiLSTM 得到的语义特征和命名实体，按照迭代训练得到的注意力权重进行分配，集中关注其中的重点部分，从而在一定程度上改善了中文问答任务的完成情况。

（3）界面设计：

针对需要实现的问答功能，问答系统界面分别设计了“选择文件”、“导入”、问答方法选择、“查询”、“帮助”、“退出”按钮，各个按钮的布局以及界面的整体设计如下图所示：

图 4-4 KBQA 实现效果 1

系统的首次使用前需要先点击“选择文件”按钮，选择系统需要使用的知识库文件，点击后：

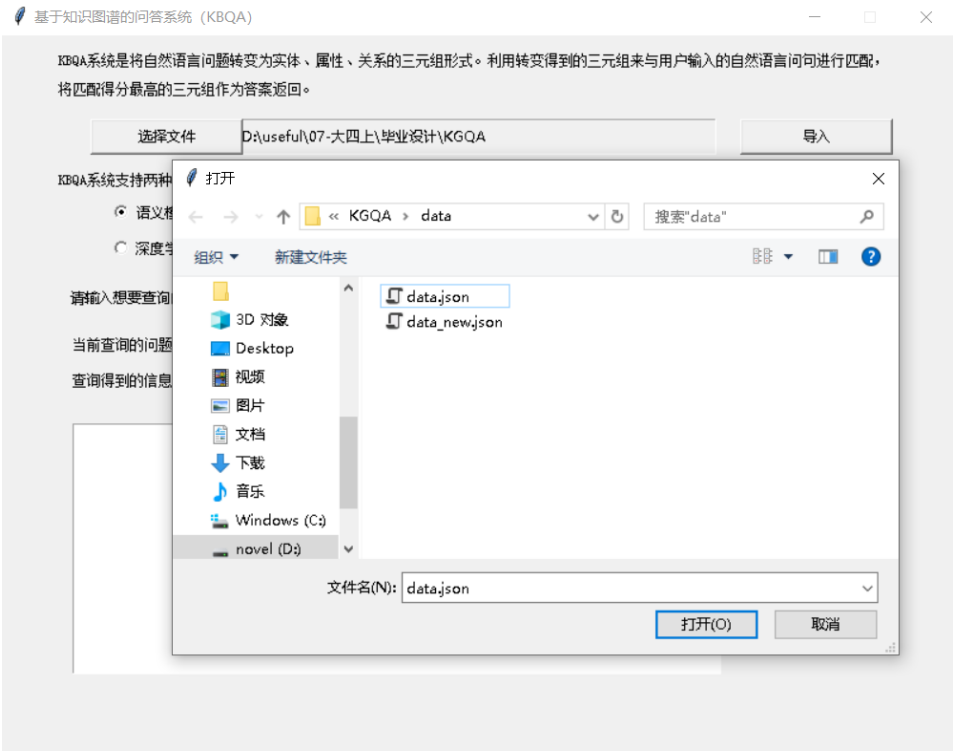

图 4-5 KBQA 实现效果 2

选中需要使用的文件，如 data.json 文件，点击打开按钮，系统将会获得文件的存储路径。之后，点击“导入”按钮，系统将会使用获得的文件存储路径对该文件进行转换，将其导入 Neo4j 图数据库中，从而可以供系统使用。

导入前的 data.json 文件形式如下：

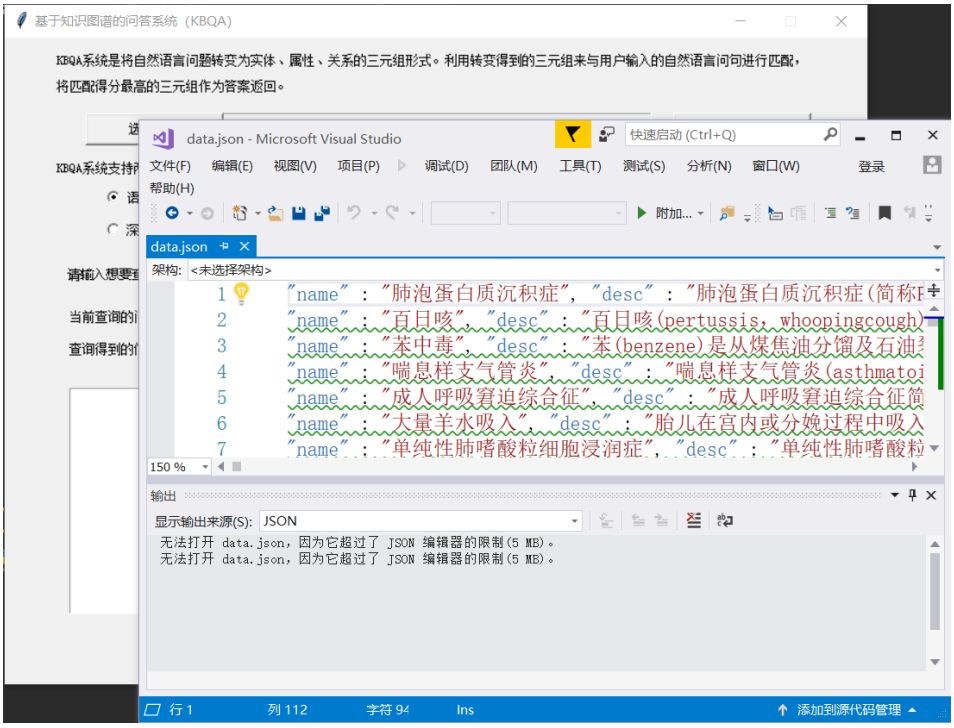

图 4-6 KBQA 实现效果 3

导入后，json 文件中的数据将以知识图谱的形式存储于 Neo4j 图数据库中，其形式如下图所示：

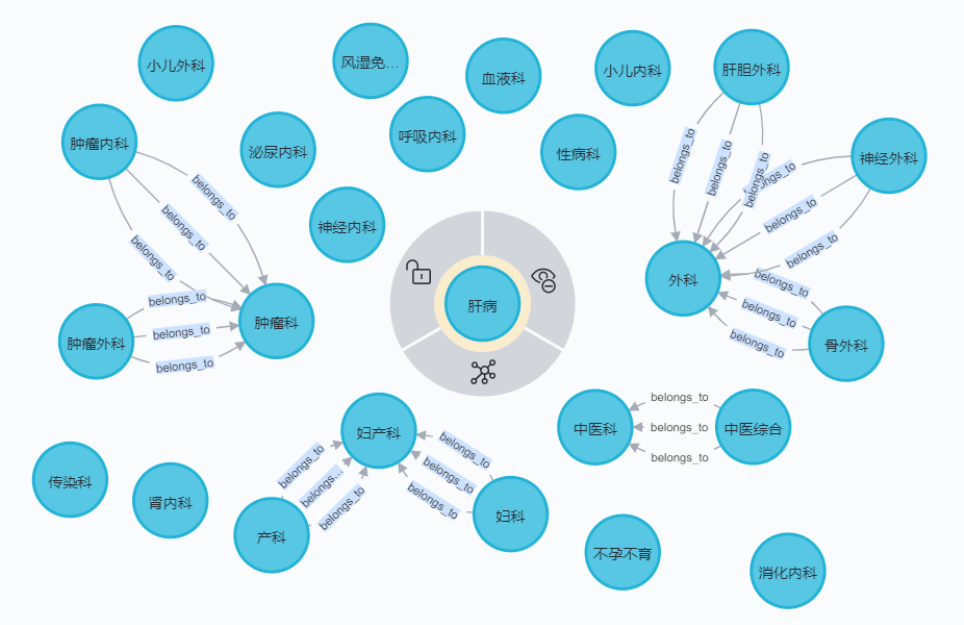

图 4-7 KBQA 实现效果 4

在使用问答系统时，首先从两种问答方法中选择一种，之后在输入栏中输入需要查询的自然语言问句，点击输入栏右侧的“查询”按钮，系统查询得到的信息将会显示在对话框下方的显示栏中。

图 4-8 KBQA 实现效果 5

图 4-9 KBQA 实现效果 6

第五章　实验总结

基于知识图谱的问答系统，除了具有能够对用户输入的自然语言问句给出精准答案的能力，还能够与传统的搜索引擎相结合，用于对传统的搜索引擎检索到的答案进行筛选，从而减少了用户对于搜索引擎返回的信息的二次筛选，极大地改善了搜索引擎的使用体验。

本研究将整个 KBQA 的工作流程划分为两个阶段：实体提取与向量映射。

整个流程在经典 KBQA 的方法上融入了深度学习方法加以改进。本研究通过改进后的 BERT-BiLSTM-CRF 模型进行问句命名实体识别。模型最大化利用 RNN 对于时序信息的提取能力，提取自然语言问句中的局域语义信息和全局语义信息，从而极大地改善了命名实体识别任务的准确率。同时，使用编辑距离来对命名实体识别任务中未准确识别的实体作进一步优化，较好地改善了实体提取任务的完成效果。

为了更为直观地展示传统KBQA方法与结合了深度学习技术的KBQA方法之间的差异，本研究在 KBQA 系统中同时实现了两种问答方法。模型在语义及词汇两个层级，对用户输入的自然语言问句以及系统知识库中筛选出的侯选属性进行编码，编码后得到二者的语义向量与词汇向量，将两个向量拼接后得到的自然语言问句的语义向量，以及知识库候选属性的语义向量。再把这两个向量映射到同一向量空间中，采用余弦距离这一指标来量化向量间相似性，最后筛选出与输入的自然语言问句最匹配的答案，将其作为结果输出。

在测试方面，本文利用公开的开放域领域知识问答数据集以及人工预先准备的数据集对模型作了测试，并构建了一个完整的基于知识图谱的问答系统。最终的实验结果证实了本研究设计并实现的基于知识图谱的问答系统的可行性。

综上所述，本文在现有的中文知识问答系统的基础之上做了更进一步的研究，将深度学习的技术与传统的中文知识问答系统加以结合，利用其能够自动对输入文本的特征进行提取的能力优化传统 KBQA 方法。实验结果也证明了这一思路的有效性。

同时，本研究仍然有一些不足之处，在后续研究之中，将会从以下环节加以改进：

第一，本研究所设计实现的深度学习模型在训练过程中，使用的词汇向量较为单一，如果使用更为优秀的词汇向量预训练方法，将有助改善模型的表现。

第二，本研究在设计实现时，系统的测试环节仅仅使用了单一通用开放领域中文 KBQA 数据集，未能较好地验证本研究所构造的模型的可移植性与泛化性，在后续的进一步研究之中，可以尝试使用更多其他数据集，对模型的可移植性与泛化性进行量化验证。

第三，由于本系统使用的数据库主要来自垂直医药网站，其中的药品名称大部分只有其成分名称，缺少相应的口语化俗名，在后续的进一步研究中，考虑在系统现有的知识库中增加相应的属性，使得系统能够对包含口语化名称的药品问题进行回答。

参考文献

[1] 李殿涛. 大数据时代人工智能在计算机网络技术中的应用[J]. 内江科技, 2021, 42(04): 114-115.

[2] Pujara J, Miao H, Getoor L, et al. Knowledge graph identification[C]//International Semantic Web Conference. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013: 542-557.

[3] 黄恒琪, 于娟, 廖晓, 等. 知识图谱研究综述[J]. 计算机系统应用, 2019, 28(06): 3-14.

[4] 王智悦, 于清, 王楠, 等. 基于知识图谱的智能问答系统研究综述[J]. 计算机工程与应用, 2020, 56(23): 1-11.

[5] 曹倩, 赵一鸣. 知识图谱的技术实现流程及相关应用[J]. 情报理论与实践, 2015, 38(12): 127-132.

[6] Weizenbaum, Joseph. ELIZA-A Computer Program For the Study of Natural Language Communication Between Man And Machine[J]. Communications of the ACM, 1983, 9(1): 36-45.

[7] Yao X, Van Durme B. Information extraction over structured data: Question answering with freebase[C]//Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. Baltimore, Maryland: Association for Computational Linguistics, 2014: 956-966.

[8] SHEN C, HUANG T, LIANG X, et al. Chinese Knowledge Base Question Answering by Attention-Based Multi-Granularity Model[J]. Information, 2018, 9(4): 98.

[9] Dong L, Wei F, Zhou M, et al. Question answering over freebase with multi-column convolutional neural networks[C]//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. Beijing, China: Association for Computational Linguistics, 2015: 260-269.

[10] Bao J, Duan N, Yan Z, et al. Constraint-Based Question Answering with Knowledge Graph[C]//Proceddings of COLING 2016, the 26th International Conference on Computational Linguistics. Osaka, Japan: The COLING 2016 Organizing Committee, 2016: 2503-2514.

[11] Jain S. Question answering over knowledge base using factual memory networks[C]//Proceedings of the NAACL student research workshop. San Diego, California:Association for Compurational Linguistics, 2016: 109-115.

[12] Chen Y, Wu L, Zaki M J. Bidirectional Attentive Memory Networks for Question Answering over Knowledge Bases[C]//Proceedings of NAACL-HL

T. Minneapolis, Minnesota: Association for Compurational Linguistics, 2019: 2913-2923.

[13] Yih W, Chang M W, He X, et al. Semantic Parsing via Staged Query Graph Generation: Question Answering with Knowledge Base[C]//Proceedings of the Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. Beijing, China: Association for Computational Linguistics, 2015: 1321-1331.

[14] Yu M, Yin W, Hsasn K S, et al. Improved Neural Relation Detection for Knowledge Base Question Answering[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Vanconver, Canada: Association for Computational Linguistics, 2017: 571-581.

[15] Tran K, Bisazza A, Monz C. Recurrent Memory Network for Language Modeling[C]//Processing of the 2016 Conference of the North American Chapter of the Assosiaction for Computional Linguistics. San Diego, California: Association for Computational Linguistics, 2016: 321-331.

[16] Xu K, Wu L, Wang Z, et al. Exploiting Rich Syntactic Information for Semantic Parsing with Graph-to-Sequence Model[C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brusselss, Belgium: Association for Computational Linguistics, 2018: 918-924.

[17] Hu S, Zou L, Zhang X. A State-transition Framework to Answer Complex Questions over Knowledge Base[C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium: Association for Computational Linguistics, 2018: 2098-2108.

[18] 周博通, 孙承杰, 林磊, 等. InsunKBQA:一个基于知识库的问答系统[J]. 智能计算机与应用, 2017, 7(005): 150-1544.

[19] Borders A, Chopra S, Weston J. Question Answering with Subgraph Embeddings[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Doha, Qatar: Association for Computational Linguistics, 2014: 615-620.

[20] 赵小虎, 赵成龙. 基于多特征语义匹配的知识库问答系统[J]. 计算机应用, 2020, 40(07): 1873-1878.

[21] 徐增林, 盛泳潘, 贺丽荣, 等. 知识图谱技术综述[J]. 电子科技大学学报, 2016, 45(04): 589-606.

[22] 陈曙东, 欧阳小叶. 命名实体识别技术综述[J]. 无线电通信技术, 2020, 046(003): 251-260.

[23] Lample G, Ballesteros M, Subramanian S, et al. Neural Architectures for Named Entity Recognition[C]//Proceedings of NAACL-HLT 2016. San

Diego: Association for Compurational Linguistics, 2016: 260-270.

[24] Huang Z， Wei X， Kai Y . Bidirectional LSTM-CRF Models for Sequ ence Tagging[J]. Computer Science, 2015.

[25] Hochreiter S, Schmidhuber J. Long Short-Term Memory[J]. Neural Comput ation, 1997, 9(8): 1735-1780.

[26] Mikolov T, Karafiat M, Burget L, et al. Recurrent neural network based la nguage model[C]//INTERSPEECH 2010, Conference of the International Sp eech Communication Association. Makuhari Chiba, Japan: September DBLP, 2010: 1045-1048.

[27] Pascanu R, Mikolov T, Bengio Y. On the difficulty of training recurrent n eural networks[C]//International conference on machine learning. PMLR. At lanta, GA, USA: JMLR, 2013: 1310-1318.

[28] 王春柳, 杨永辉, 邓霏, 等. 文本相似度计算方法研究综述[J]. 情报科学, 20 19, 37(03): 158-168.

[29] Dai Z, Li L, Xu W. CFO: Conditional Focused Neural Question Answerin g with Large-scale Knowledge Bases[C]//In Proceedings of the 54th Annua l Meeting of the Association for Computational Linguistics. Berlin, Germa ny: Association for Computational Linguistics, 2016: 800-810.

[30] He K, Zhang X, Ren S, et al. Deep residual learning for image recognitio n[c]//Proceedings of the IEEE conference on computer vision and pattern r ecognition.Los Alamitos, CA, USA: IEEE Computer Society, 2016: 770-77 8.

[31]陈子睿,王鑫,王林,徐大为,贾勇哲.开放领域知识图谱问答研究综述[J].计算机科学与探索,2021,15(10):1843-1869.

[32]Sui Y. Question answering system based on tourism knowledge graph[C]//Jo urnal of Physics: Conference Series. IOP Publishing, 2021, 1883(1): 01206 4.

[33]Kai Ding,Hongqi Han,Linna Li,Menglin Yi. Research on Question Answerin g System for COVID-19 Based on Knowledge Graph[C]//第 40 届中国控制会议论文集（8）.[出版者不详],2021:550-555.DOI:10.26914/c.cnkihy.2021.02 8719.

[34] Graves A, Schmidhuber J. Framewise phoneme classification with bidirecti onal LSTM and other neural network architectures[J]. Neural Networks, 20 05, 18(5): 602-610.

[35]Mnih V, Heess N, Graves A, et al. Recurrent models of visual attention[C] //Proceddings of the 27th International Conference on Neural Information P rocessing Systems. Cambridge, MA, USA: MIT Press, 2014: 2204-2212.

[36]Phuc Do,Truong H. V. Phan. Developing a BERT based triple classification model using knowledge graph embedding for question answering system [J]. Applied Intelligence,2021(prepublish):

[37]Truong H. V Phan,Phuc Do. BERT+vnKG: Using Deep Learning and Kno wledge Graph to Improve Vietnamese Question Answering System[J]. Inter national Journal of Advanced Computer Science and Applications (IJACS A),2020,11(7):

[38] Kai Ding,Hongqi Han,Linna Li,Menglin Yi. Research on Question Answeri ng System for COVID-19 Based on Knowledge Graph[C]//第 40 届中国控制会议论文集（8）.[出版者不详],2021:550-555.DOI:10.26914/c.cnkihy.2021.028719.

[39]Jiang Z, Chi C, Zhan Y. Research on Medical Question Answering System Based on Knowledge Graph[J]. IEEE Access, 2021, 9: 21094-21101.

致　谢

时光荏苒，岁月如梭，转眼之间四年的大学生活已经接近尾声。在这段并不太短的时间里，我遇到了学识渊博的老师们，结识了志同道合的朋友、学长、学姐和同学，经历了很多事情，也学到了很多知识与经验。

本次研究，也是为大学四年的学习生活画上一个句号。

在完成论文的撰写过程中，我最感谢的是我的导师，饶国政教授。饶老师的专业知识功底十分深厚，治学严谨，为我的课题研究的顺利完成提供了很大的引导与帮助。很感谢饶老师为我提供的学术上的专业指导与帮助，我将在未来的学习生活之中，学习饶老师的严谨的治学态度，一丝不苟地对待科研。

此外，我要感谢我的家人对我的莫大支持，感谢所有帮助过我的老师与同学，感谢所有审核与答辩的老师，感谢你们在百忙之中拨冗审阅我的论文。你们辛苦了！