学校代码 10459
学号或申请 201512172165
密　　级

郑州大学

硕士学位论文

问句分类方法及其在问答系统中的应用研究

作　者　姓　名：张倩
导　师　姓　名：穆玲玲 副教授
学　科　门　类：工　学
专　业　名　称：软件工程
培　养　院　系：信息工程学院
完　成　时　间：2018 年 5 月

A thesis submitted to

Zhengzhou University

for the degree of Master

Question classification method and its application in question answering system

By Qian Zhang

Supervisor:A/Pro.Lingling Mu

Software Engineering

School of Information Engineering

May 2018

学位论文原创性声明

本人郑重声明：所呈交的学位论文，是本人在导师的指导下，独立进行研究所取得的成果。除文中已经注明引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的科研成果。对本文的研究作出重要贡献的个人和集体，均已在文中以明确方式标明。本声明的法律责任由本人承担。

学位论文作者：张倩　　日期：2018年5月27日

学位论文使用授权声明

本人在导师指导下完成的论文及相关的职务作品，知识产权归属郑州大学。根据郑州大学有关保留、使用学位论文的规定，同意学校保留或向国家有关部门或机构送交论文的复印件和电子版，允许论文被查阅和借阅；本人授权郑州大学可以将本学位论文的全部或部分编入有关数据库进行检索，可以采用影印、缩印或者其他复制手段保存论文和汇编本学位论文。本人离校后发表、使用学位论文或与该学位论文直接相关的学术论文或成果时，第一署名单位仍然为郑州大学。保密论文在解密后应遵守此规定。

学位论文作者：张倩　　日期：2018年5月27日

摘要

传统的信息检索系统均使用关键词组合作为系统输入，忽略了问句语义的多样性和语言结构的分析。问答系统能够接受用户以自然语言形式描述的问题，并能从大量异构的数据中查找或推断出用户问题的答案，提高用户查询效率。因此问答系统成为信息检索技术向人性化、智能化方向发展的一种必然趋向。

问句分析的目的是明确用户意图，有效地定位到正确答案。因此，问句分析是问答系统的核心技术之一，而问句分类是问句分析的重要组成部分。

在深入学习了目前中文问句分类及问答系统相关研究方法的基础上，本文提出了基于最大熵模型和双向长短期记忆人工神经网络（Bi-LSTM）模型的问句分类方法，具体研究工作如下：

（1）研究了基于最大熵模型的问句分类方法。该方法把句法分析和词向量等语义知识运用到问句表示中，研究了问句的词汇特征、句法特征和词向量特征对问句粗分类准确性的影响，实验结果表明，相对于其他特征，词向量特征对问句粗分类取得了较好的效果，准确率达到 88.75%。

（2）研究了基于 Bi-LSTM 的问句分类方法。基于最大熵模型的问句分类方法需要人工提取问句的特征，带有一定的主观性。而基于 Bi-LSTM 的问句分类方法能够自主地学习问句的句法和语义特征，避免了人为因素带来的干扰。在分类模型中，本文使用了词语、词性和位置特征，并将这三种特征向量融合得到的词嵌入作为模型的输入，然后将输出结果通过最大池化层（Max Pooling）和 Softmax 层来完成问句特征提取和问句粗分类的工作。实验结果表明，该方法在粗粒度分类上准确率达到 92.38%。

（3）研究了问句分类在知识库问答系统中的应用。本文借助问句分类特征，再结合相似度、编辑距离和共现特征，利用 Ranking SVM 算法对候选答案进行排序。在 NLPCC2016 开放域知识库问答系统的评测任务的数据集上进行实验，结果表明，将问句分类应用到知识库问答系统的答案排序中，有助于提高答案识别的准确率，其准确率达到 74.49%，召回率达到 83.20%，平均 F1 值达到 76.13%。

关键词：问句分析；问句分类；最大熵模型；Bi-LSTM；Ranking SVM；知识库问答系统；

Abstract

Traditional information retrieval systems use keyword combinations as the input of system, ignoring questions' semantic diversity and language structure analyzing. The question answering system can accept natural language questions, which can find or infer answer to users' questions from a large number of heterogeneous data and improve users' query efficiency. Therefore, question answering system has become an inevitable trend for information retrieval technology to the direction of humanization and intelligence.

The purpose of question analysis is to clarify the user's intention, and effectively locate the right answer. Therefore, the question analysis is one of the core techniques of the question answering system, question classification is an important part of the question analysis.

After deeply studying the current research methods of Chinese question classification and question answering system, this thesis proposes a question classification method based on the maximum entropy model and bidirectional LSTM (Bi-LSTM) networks model. The specific research work is as follows:

(1) We proposed a question classification method based on Maximum Entropy Model. This method applies semantic knowledge, such as syntactic structure and word vector, to represent question, and studies the impacts of lexical feature, syntactic features, and word vector features on the accuracy of coarse classification of questions.Experimental results showed that compared with other features, the word vector feature has a good effect on coarse classification of questions, and the accuracy rate reaches 88.75%.

(2)This thesis proposed a question classification method based on Bi-LSTM. The question classification method based on the maximum entropy model needs manually extracting the features of the question, with a certain degree of subjectivity. The question classification method based on Bi-LSTM can autonomously learn the syntax and semantic features of question and avoid the interference caused by human factors. In the classification model, this thesis uses words, parts of speech and words

location as features. The word embedding obtained by the fusion of the three feature vectors is used as the input of the model, the output is used to obtain question feature and coarse-grained classification through Max Pooling layer and Softmax layer. The experimental results showed that the accuracy is 92.38% on coarse classification.

(3) The application of the question classification in the knowledge base question answering system. This thesis uses the Ranking SVM algorithm to sort the candidate answers by using the question classification, the similarity, the editing distance and the co-occurrence feature. Experiments are conducted on the data set of the NLPCC2016 open domain knowledge base question answering system. The results showed that applying the question classification to the answer sorting of the knowledge base question answering system helps improving the accuracy of answer recognition. The accuracy rate reaches 74.49%, the recall rate reaches 83.20%, and the average F1 reaches 76.13%.

Key words : Question parsing; Question classification;Maximum Entropy Model; Bi-LSTM;Ranking SVM; Knowledge base question answering system;

目录

图表目录

图目录

表目录

1 引言

随着中文信息处理技术的迅猛发展，人们快速获取准确、简洁信息的愿望越来越迫切，问答系统应运而生，逐渐受到国内外专家的高度重视和关注。问句分析是中文智能化问答系统的基石，而问句分类则是其中的重要组成部分。本章讲述问答系统的研究背景和意义以及问句分类在问答系统中的作用。最后叙述本文主要内容并介绍章节布局。

1.1 研究背景和意义

当今社会是一个信息呈现爆炸式增长的时代，随着计算机和互联网的迅猛发展，社会信息量也层出不穷，网上的信息资源瞬间丰富得超乎人们的想象，真正让你做到足不出户就可以随时随地知道天下事。然而，互联网信息的急速膨胀为人们带来海量信息的同时，一些挑战也随之而来。面对海量、碎片化的信息，如何从错综复杂的信息中高效准确地获得目标信息成为了当今社会的一大难点。传统的搜索引擎通过关键词组合从互联网上浏览和检索人们所需要的相关信息，然后按照一定的排列顺序返回相关文件集合或文件链接集合。但是，传统的搜索引擎存在一些不足：首先，以关键词组合为基础的索引、匹配算法没有触及深层的语义，因此很难进一步提高效果。其次，检索结果的文件集合或链接集合往往并不是用户想要的最终结果。最后，传统的搜索引擎以关键词组合作为输入，用户往往还得从大量无关的文档或链接集合中找到自己想要的信息，这样会耗费大量的时间，且得到的答案也非常地繁琐复杂。

因此，如何精准地定位正确答案并快速、简洁地作出回答成为当今用户的迫切需求。于是人们逐渐地把注意力转移到问答系统上来，问答系统应运而生。与传统搜索引擎不同的是，问答系统可以直接接受用户以自然语言的形式提出的问题，并且把答案直接提交给用户，节省了大量的时间，而且符合人性化的设计理念。另外，它能更好地满足用户的需求，更准确地理解用户的意图，能更快地定位到用户所需要的答案，免去了用户从繁琐的文档集合中进一步筛选的麻烦。

中文问答系统的体系结构主要包含三个部分：问句分析，信息检索和答案

抽取[1]。问句分析是问答系统的基石，而此部分中较为重要的是问句分类。问句分类就是在确定的分类标准下，根据答案的特点或问句的语义信息确定问句所关联的类型。对问句进行类别标注，不仅可以有效地缩减候选答案的查找空间，而且还能够减少定位答案正确位置的时间。另外，问句的类型信息往往也是答案的类型，这可以直接决定答案的抽取策略，提升查找答案的准确率。问句分类准确率影响着整个问答系统的质量和性能。因此，研究问句分类对提高问答系统的质量和性能有着重要的作用。

1.2 论文主要研究工作

由于问句分析是问答系统的基础，其中较为重要的部分是问句分类，其结果的好坏对后续答案的抽取具有重要的作用。因此，如何提高问句粗分类的准确率成为本文首要研究的工作。本文采用传统的基于统计机器学习的方法和目前最热的深度学习方法对问句进行粗粒度分类。然后，本文将问句粗粒度分类的结果应用到知识库问答系统中作为最后研究的主要工作。研究的主要内容如下：

（1） 基于最大熵模型的问句分类方法。此方法提取四种特征分别对问句进行粗分类，它们是：词袋特征、词和词性特征、句法特征、词向量特征。词袋特征是对生语料进行分词后得到的词序列；词和词性特征是对生语料进行分词和词性标注后得到的词和词性的混合序列；问句的句法特征是根据依存关系得到的五元组词序列；词向量特征是根据预训练好的每个词向量得到问句的向量序列。本章根据以上四种特征分别利用最大熵模型对问句进行粗粒度分类，实验结果表明，与其他三种特征相比，利用词向量特征对问句粗粒度分类达到较好的效果，其准确率达到 88.75%。

（2） 基于双向长短期记忆人工神经网络（Bi-LSTM）模型的问句分类方法。此方法首先对生语料进行预处理，然后提取词、词性和词语位置特征，并将其融合生成词嵌入。最后将生成的词嵌入作为 Bi-LSTM 模型的输入，且在最大池化层（Max Pooling）提取问句的特征，在 Softmax 层对问句进行粗粒度分类，准确率达到 92.38%。

（3） 问句分类在知识库问答系统中的应用。在知识库问答系统的答案排序阶段，将问句分类作为一个特征，并结合相似度、共现、编辑距离特征，采用

Ranking SVM 算法对候选答案三元组进行排序，得分最高的成为问句的正确答案。实验结果表明，提高问句分类的准确率有助于提高知识库问答系统中答案识别的准确率，其准确率达到 74.49%，召回率达到 83.20%，平均 F1 值达到 76.13%。

1.3 论文组织结构

根据本文对中文问句分类方法及其在知识库问答系统中的应用的研究，将本文分为六章，各章节的安排如下：

第一章，引言。介绍了问答系统的研究背景和意义，以及问句分类对问答系统的作用。接下来，介绍论文的主要研究工作和章节安排。

第二章，相关工作。问句分析作为问答系统的基石，它对全面正确理解问句的语义表示具有重要的作用。问句分析包括问句分类、关键词提取和关键词扩展三个部分，其中较为重要的部分是问句分类。问句分类可以缩减候选答案的搜索空间，并且减少查找正确答案的时间，对抽取答案的策略具有一定的指导意义。本章首先介绍了国内外问句分类的相关研究，然后介绍了知识库问答系统的相关研究。

第三章，基于最大熵模型的问句分类。好的问句分类结果能在一定程度上提升问答系统的性能。问句分类的目标就是为每个问句分配一个类别，该类别也代表了问句期望的答案的类型。本章提出了基于最大熵的问句分类模型，该模型分别提取词袋特征、词和词性特征、句法特征、词向量特征，分别利用最大熵模型对问句进行粗粒度分类。文中还分析了四种特征对问句粗分类的影响。

第四章，基于 Bi-LSTM 的问句分类。随着深度学习在自然语言处理任务中的广泛应用，本章提出了基于 Bi-LSTM 的问句分类方法，该方法能够自主地学习问句的特征，有助于全面理解问句的语义，相比传统的问句分类方法具有一定的优势。本章融合词向量、词性和词语位置特征生成词语的嵌入表示，通过 Bi-LSTM 的隐藏层提取问句的分布式特征，通过最大池化层生成问句的表示，并在 Softmax 层对问句进行分类。文中将此方法的实验结果与传统的方法作了对比。

第五章，问句分类在知识库问答中的应用。知识库是以三元组的形式表示的。因此为了匹配知识库中的答案，首先要识别出问句中的话题，然后再将问

句结构化。本章将问句中话题的识别看成二分类问题，提取前词、后词、逆文档频率作为分类特征，并利用随机森林算法来识别问句中的话题，根据识别出的话题从知识库中抽取每个问句的候选答案三元组。最后，提取共现、编辑距离和相似度特征，再结合问句分类特征，利用 Ranking SVM 算法对每个候选答案排序，得分最高的作为最终的正确答案。

第六章，总结和展望。本章主要是对本文的研究工作进行概括和总结。并展望了下一步将问句的句法和语义特征应用到 Bi-LSTM 模型中，进一步提高问句的分类性能。另外，还展望了下一步采用深度学习的方法进一步提高问答系统的性能。

2 相关工作

2.1 问句分类方法的相关研究

问答系统能用准确、简洁的自然语言回答用户提出的问题，一般包括问句分析、信息检索和答案抽取三个模块。问句分析作为问答系统的第一步，可分为三个部分：问句分类、关键词扩展和关键词提取[1]。其中较为重要的部分是问句分类。问句分类是在给定问句的情况下，将问句映射到预定义的 K 个类别之一，此 K 个类别表示对所寻求的答案的语义约束。问答系统中的问句分类模块主要有两个用途：首先，它提供了对答案类型的约束，允许进一步处理精确定位和验证答案。第二，它提供了答案选择策略的信息，这些策略不是统一的，而是针对特定答案类型制定的。例如：问句“*初唐四杰中有谁*”，针对该问句，只要知道该问句的目标是一个“人”，就能较大地减少可能答案的搜索空间。

传统的问句分类方法有基于规则的方法和基于统计机器学习的方法。其中，基于统计机器学习的问句分类方法占主导地位。近年来，深度学习在自然语言处理（Natural Language Processing，NLP）领域中得到了广泛的应用，一些研究人员开始利用深度学习的方法对问句进行分类，因此，基于深度学习的问句分类方法开始兴起。以上这些方法都是为了给问句分配一个对应的类别，然而要对问句进行类别标注，首先要了解问句有哪些类型，而问句的类型是由采用的分类体系决定的。因此，问句分类体系是研究各种问句分类方法的前提和基础。

2.1.1 问句分类体系

早期的工作已经提出了各种各样的问句分类体系。1986 年 Wendy Lehner[2] 提出了一个概念分类体系，该体系定义了大约 13 个概念类，它们分别是：“因果的前因（casual antecedent）”、“因果相随（causal consequent）”、“目标导向（goal orientation）”、“验证（verfication）”、“启用（enablement）”、“析取（disjunctive）”等。2002 年 Kuo 和 Lin[3]将 ATIS 语料库中包含的 5513 个问句分为八类：“leg”、“selection”、“price”、“description”、“constraint”、“time”、“location”和“other”，并将其应用到铁路信息服务问答系统中，

使问答系统的准确率提高了 5 个百分点。然而，这些分类体系不适用于事实问题，它并不能满足人们对问句进行语义分类的要求。因此，2002 年 Li 和 Roth[4] 等提出了一个分层分类体系，该体系定义了 6 个粗类和 50 个细类，粗类分别为：“ABBREVIATION”、“ENTITY”、“DESCRIPTION”、“HUMAN”、“LOCATION”和“NUMERIC VALUE”。它满足了人们对问句进行语义分类的要求，因此该分类体系被广泛应用到研究问句分类的方法中。2003 年 Dell Zhang[5]将该分层分类体系应用到 USC[6]、UIUC[4]、TREC[7][8][9]提供的数据集上，对其进行人工标注类别。然后将标注好类别的数据集随机分成五部分作为训练集，剩下的部分数据集作为测试集，利用 SVM 方法对问句进行自动分类。2007 年 Dan Shen[10] 等为了识别 TREC02-05 提供的事实问句期待的答案类型，使用与 Li 和 Roth 相同的问句分类体系来决定答案的类型。2008 年 Zhiheng Huang[11]等根据 Li 和 Roth 提出的问句分类体系人工标注 UIUC 问句分类数据集[1]作为训练集和测试集，利用 SVM 和最大熵算法对问句进行自动分类。2015 年 Chalabi 和 Rays[12] 等根据 Li 和 Roth 提出的问句分类体系对阿拉伯问句进行分类，使阿拉伯语言问答系统的召回率达到了 0.93.

虽然 Li 和 Roth 提出的问句分类体系为一些学者研究问句的自动分类奠定了良好的基础，但是该体系是针对英文问句集提出的，为了满足中文问答系统研究的迫切需求，一些学者开始致力于中文问句分类的研究。2005 年余正涛[13] 等根据问句实际分布情况提出了一个两层的中文问句分类体系，该体系将问句集分为 6 个粗类，分别为：“缩写”、“实体”、“描述”、“人”、“位置”、“数量”。然后，利用 SVM 进行汉语问句分类获得较好的效果。2016 年董才正[14]等根据问句自身的语义信息和问答社区中问句类型分布的特点，提出了一种面向问答社区的粗粒度分类体系，该体系包含 7 种问句类型，分别为：“定义”、“事实”、“过程”、“原因”、“观点”、“是非”、“描述”。然后，根据该分类体系手工标注中文问答社区知乎的 4103 个问句作为训练和测试集，利用 SVM 对问句进行自动分类。2006 年文勖[15]等根据现实世界对事物划分的特点，在国外一些已有的问句分类体系基础上，定义了中文问句分类的分类体系，该体系包含 7 个大类，分别为：“人物”、“地点”、“数字”、“时间”、“实体”、“描述”、“未知”，并根据该分类体系人工标注 3300 个中文问句集作为训练和测试集，然后利用贝叶斯算法对问句进行自动分类。由于

[1]http://cogcomp.cs.illinois.edu/Data/QA/QC

该体系对今后在开放域问答系统领域的研究工作有着重要的意义，因此，它被大多数学者所采纳，广泛应用到各种研究中文问句分类方法的数据集上。

2.1.2 基于规则的问句分类方法

基于规则的问句分类方法主要是通过提取各种类型问句中的疑问词和相关词的特征规则，通过规则来判断问句的类型。2002 年 Bernardo Magnini[16]等为了扩大可能的问句类型的覆盖范围并优化答案类型分布，在新版的 DIOGENE 问答系统的答案类型识别模块中，采用一组约 340 条规则来检查输入问句的不同特征，并根据这些特征识别问句的类型。2003 年 Hui Yang[17]等在抽取与问句类别匹配的答案前，首先对文档检索和答案排序阶段获得的前 k 个位置的问句进行细粒度的命名实体识别标记，并结合词汇和语义特征采用基于规则的算法对问句进行命名实体识别标记。测试结果的准确率达到了 90%以上。2004 年樊孝忠[18]等定义两种规则：基于问点块的组成规则和基于语义块的问句句型模板规则，对银行领域自动问答系统（BAQS）的问句进行分类。2009 年辛霄[19]等考虑到面向真实环境的问答系统的问句相对复杂，对非事实类的问句，主要是结合金融领域问句自身的特点，以及问句的语义来对问句进行分类。2010 年贾君枝[20]等针对农民问答系统中问句自身的特点，建立了针对非正式疑问词和无疑问词时的“特殊规则表”对问句进行分类。2011 年任梦菲[21]等将股票领域问答系统定位于三大类问题：基本概念类、个股信息类、投资家类。然后基于这个分类标准提出了更加完善的适用于限定域的分类规则，大大提高了问答系统的实用性。

基于规则的问句分类方法实现起来相对比较容易，而且对问句自动分类的速度较快，不需要大量的训练数据。但是为了得到较好的实验结果，此方法往往需要大量的规则，而且这些规则需要人工书写，当问句分类体系和问句表示形式发生改变时，这些规则也需要作出相应的调整，灵活度不高，扩展性不强。当训练语料较大时，这将是一个耗时的工作量。另外，这种方法需要构建一个非常完整而且准确的规则库，而规则库在开放域问句的处理上适用性不强。

2.1.3 基于统计机器学习的问句分类方法

随着统计方法的日益普及，机器学习在问句分类任务中发挥着越来越重要的作用。与传统的基于规则的问句分类方法不同的是：该方法可以借助数千个

或更多的问句特征自动构建一个高性能的问句分类算法。如果给定更多的训练数据训练分类算法，它的性能通常会提高。另外，训练分类算法可以很容易地适应新的领域，因此它比基于规则的分类算法更灵活，具有更强的扩展性。

基于统计机器学习的问句分类方法主要集中在三个方面进行研究：特征工程、特征选择和采用不同机器学习模型进行分类。首先，在特征工程方面，词袋特征是最常用的特征，除此之外，还有一些复杂的特征，例如：词性、名词短语[22]和树核函数[23]特征等。特征选择的目标就是找到最有效的特征来提高问句分类的性能，最广泛使用的特征选择方法是词频。另外，也会经常使用一些高效的方法来选择有效的特征，例如：信息熵、互信息[24]或 L1 正则化[25]等。问句分类中常用的机器学习模型有贝叶斯模型(Bayes)[15][26][27]、最大熵模型(EM)[11][28]、支持向量机（SVM）模型[5][29]。李鑫[29]等利用统计量选择 wordnet 中的上位概念对问句中的词汇进行选择型扩展，在 UIUC 数据集[1]上进行粗粒度分类，准确率达到 91.60%。Dell Zhang[5]等提出了一种树核（tree kernel）函数，它能使 SVM 充分利用问句的句法结构进行分类，在 TREC QA track 数据集[1]上进行粗粒度分类，准确率达到 90.0%。张宇[26]等利用词与词之间的无关性简化问句分类，在哈尔滨工业大学信息检索实验室提供的中文问句集上进行粗粒度分类，准确率达到 72.4%。田卫东[27]等提出了用自学习来抽取规则并结合贝叶斯对问句进行分类的方法，在哈尔滨工业大学信息检索实验室提供的中文问句集上进行粗粒度分类，准确率达到 84%。文勖[15]等提取问句的主干和疑问词及其附属成分作为分类特征，在哈尔滨工业大学信息检索实验室和中科院自动化研究所实验室提供的中文问句集上进行粗粒度分类，准确率达到 86.62%。Huang Z[11]等提取问句的焦点词作为特征，并借助 WordNet[2]扩充焦点词的语义特征，得到焦点词的上位词特征。然后，利用最大熵模型在 UIUC 数据集[1]上进行粗粒度分类，准确率达到 89.0%。孙景广[28]等选取问句的疑问词、疑问意向词和意向词在知网[3]中的首义原作为分类的特征，在哈工大检索实验室和中科院自动化实验室提供的中文问句集上进行粗粒度分类，准确率达到 92.18%。

为了获得问句特征，基于统计的机器学习方法一般需要先对问句完成词性标注、句法分析和语义分析等自然语言处理（Natural Language Processing，NLP）任务，这些 NLP 任务的准确率将对问句分类准确率产生较大的影响。另外，人

[2]http://wordnet.princeton.edu/man/wnstats.7WN
[3]http://www.keenage.com/zhiwang/c_zhiwang.html

工选择特征具有一定的主观性，且需要消耗大量的人力资源。该方法采用的机器学习方法都存在数据稀疏性的问题。

2.1.4 基于深度学习的问句分类方法

近几年，深度学习模型在计算机视觉[30]、语音识别[31]和 NLP 等领域已经取得了显著的成果。预训练的词向量和深度神经网络的快速发展给各种各样的自然语言处理任务带来了新的启发。词向量是词语的分布式表示且可以大大降低数据的稀疏性问题。Mikolov,Yih 和 Zweig[32]证明预训练的词向量能够捕捉到有效的句法和语义规则。词向量中的词是通过隐藏层从稀疏的 1-V 编码（这里 V 是词汇大小）映射到一个较低维向量空间，这对于将词的句法和语义特征编码到词向量中的特征抽取来说是非常重要的。在这样密集的表示中，语义上接近的词，将它们映射到低维向量空间上的余弦距离也是接近的，这样就很好地解决了传统问句分类方法的数据稀疏性问题。

与传统的机器学习方法不同的是，深度学习一方面不需要人工抽取问句的特征，降低了人力成本和时间成本。它能够自动地获取基本的特征，然后将这些基本特征组合成复杂的特征，最后训练模型来捕捉提取的问句特征和问句类别之间的语义关系。另一方面，它不使用传统的 N 元语言模型，因为训练模型的深度能够充分利用词序特征，例如卷积神经网络（CNN）模型中过滤器的大小可以被看作类似 N 元语言模型的形式，循环神经网络（RNN）能够捕捉到更长距离的词语信息，这可以通过权重矩阵反映出来。目前最常用的进行问句分类的深度学习模型有 CNN 模型[33][34][35]、长短期记忆网络（LSTM）模型[36]、Bi-LSTM 模型[37][38]。Xiao[33]等提出了多任务卷积神经网络进行法律问题分类，该模型共享粗粒度分类和细粒度分类的上下文信息。首先，将预训练的词向量作为 CNN 的输入进行粗粒度分类，然后把粗粒度分类的输出作为细粒度分类的输入。在中文法律问句数据集[39]上进行实验，粗粒度分类的准确率为 97.49%，细粒度分类的准确率为 92.14%。Yoon Kim[34]等针对多领域分类任务，采用 CNN 模型的四种变体来做实验，它们分别是：CNN-rand、CNN-static、CNN-non-static 和 CNN-multichannel。其中，CNN-rand 模型是基本的模型，它的输入是随机初始化所有词的向量且这些词向量会随着训练过程被修改。CNN-static 模型的输入是采用 word2vec[40]预训练好的词向量，且这些词向量在训练过程中不会被修改，训练时更新的知识模型的其他参数。CNN-non-static

模型的输入仍然是 word2vec 预训练好的词向量，但是这些词向量会随着不同的任务被微调。CNN-multichannel 模型采用两个通道来处理预训练好的词向量的两个集合，能够保证词向量的一个集合在微调的时候，其他的保持不变。在 MR[4]、SST-1[5]、SST-2[5]、Subj[41]、TREC[1]、CR[6]、MPQA[7]数据集上进行实验，实验结果表明以上四种 CNN 模型提升了七个分类任务中的四个任务（包括情感分析和问句分类）的性能。张栋[35]等将大量未标注的问句与答案样本参与到问句与答案联合学习词向量表示中，然后，将已标注的问句以词向量形式表示作为训练集，采用 CNN 建立问句分类模型。在 360 问答社区[8]上进行实验，实验结果表明该方法的分类性能明显优于传统的半监督学习方法。李超[36]等提出混合 LSTM 和 CNN 的学习框架自主学习问句特征，在哈工大、NLPCC2015 QA 和复旦大学提供的中文问句集上进行实验，准确率达到了 93.08%。周鑫鹏[37]等融合多种句子特征，使用 Bi-LSTM 模型在 UIUC 数据集上进行实验，粗粒度分类准确率达到 94.0%。徐建[38]等提出了融合双语料特征的 Bi-LSTM 问句分类方法，在 360 问答社区[8]上进行实验，实验结果表明该方法有效提高问句分类的性能。

2.2 知识库问答系统的相关研究

知识库问答系统是利用现有的知识库回答自然语言提出的问题。大多数知识库是结构化数据库，例如：DBPpedia[42]、Freebase[43]、Yago2[44]，它们通常是以三元组形式存储的，即（实体 1，关系，实体 2）。构建知识库的目标是将文本组织成以实体为节点，实体之间的语义关系为边的图结构。问答系统在构建好的知识库中查找、推理与用户问题的语义相匹配的，知识库问答应运而生。常用的解决知识库问答系统问题的方法有：基于语义分析的知识库问答、基于信息抽取的知识库问答、基于深度学习的知识库问答。

2.2.1 基于语义分析的知识库问答

基于语义分析的知识库问答首先通过语义解析器将用户问题转成逻辑形式，然后，将解析的结果生成结构化查询来搜索知识库，并获得答案。其中最

[4]https://www.cs.cornell.edu/people/pabo/movie-review-data/
[5]http://nlp.stanford.edu/sentiment/Data
[6]http://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html
[7]http://www.cs.pitt.edu/mpqa/
[8]http://wenda.so.com

重要的一步是把问句映射成预定义的逻辑形式，例如组合范畴语法（CCG）[45]和基于依赖关系的组合语义（DCS）[46]。有些基于语义分析的系统需要人工标注逻辑形式来训练解析器，而这些标注往往是非常昂贵的。因此最近的工作主要使用弱监督（问题和答案对）方法来有效训练语义分析器。Liang[47]等依据基于依赖关系的组合语义学的特性，将问句解析成树结构的逻辑形式。Kwiatkowski[48]等针对问句中本体不匹配问题引入一种新的语义解析方法。解析器从问题-答案对中学习，使用概率 CCG 构建语言驱动的逻辑形式语义表示。另外，为了调整每个目标本体的输出逻辑形式，解析器中还使用了一个本体匹配模型。Berant 和 Liang[49]等利用关联模型和向量空间模型从一组能够规范实现自然语言的候选逻辑形式中选取出一个最好的逻辑形式。这些方法在不使用逻辑形式的情况下达到了可比的结果。但是，一些方法还是要依赖词汇触发器或者人工定义特征，这限制了它们的领域和可扩展性。

2.2.2 基于信息抽取的知识库问答

基于信息抽取的方法从知识库中抽取一组候选答案，根据从问句中提取的特征，采用机器学习算法对这组候选答案排序。Lai Yuxuan[50]等提取候选答案中的谓词得分、实体长度、答案模板出现次数的特征，采用 Ranking SVM 对候选答案三元组排序。Yao Xuchen[51]等提取问句中的依存关系、疑问词、问句焦点词、问句动词、问句的话题词特征，然后将其转化为问句图，结合知识库中的话题图，寻找正确答案。Bast[52]等使用三种预定义的模板生成候选答案，然后提取重叠词的数量、派生词、词向量余弦相似度作为特征，采用机器学习算法对候选答案排序。

2.2.3 基于深度学习的知识库问答

近年来，随着深度学习在自然语言处理领域的快速发展，一些研究者开始利用深度学习解决知识库问答问题。这些方法的核心就是把用户提出的问句和知识库中的资源都映射到同一个连续的低维向量中，将问句和答案都表示成一个向量。Li Dong[53]等从答案路径、答案上下文和答案类型三个方面采用多层卷积神经网络来学习问句的表示。与此同时，它们还联合训练了知识库中的实体和关系的低维向量，最后利用问句答案对训练候选答案的排序模型。Bordes[54]等将问句及知识库中的三元组转化成向量，利用向量间的余弦相似度找出最有

可能的答案三元组。Yih[55]等把知识库问答分三步完成：首先，利用 CNN 从问句中查找与知识库中实体对应的实体；然后，利用 CNN 从问句中找到与知识库中的语义关系对应的信息。最后，根据找到的实体和关系，从知识库中找到其指向的答案实体。从以上的工作来看，基于深度学习的知识库问答需要从多个角度对问句的语义进行分析，包括问句语义与知识库中实体的匹配度以及与知识库中关系的匹配度。

2.3 本章总结

本章主要介绍了问句分类方法和知识库问答系统的相关研究工作。第一节介绍了问句分类方法的相关研究。问句分类体系是问句分类的基础，本章首先介绍了问句分类体系的相关研究，然后介绍问句分类的三种方法的研究工作及其优缺点。第二节介绍了知识库问答系统的相关研究。主要是对解决知识库问答系统的问题的三种方法的相关工作进行介绍。

3 基于最大熵模型的问句分类

基于最大熵模型的问句分类方法分别提取词袋、词和词性、句法、词向量四种特征，分别利用最大熵模型对问句进行粗粒度分类。本章提出了一种基于最大熵的问句分类模型。

3.1 最大熵模型

最大熵提供了从训练数据中估计概率分布的合理方法。其基本思想是：对未知信息不做任何假设，即当模型满足训练数据中的已知约束条件，且没有其他信息可利用时，它所估计的概率分布应该尽可能地接近均匀分布，即它所具有的熵最大。

根据 Nigam et al. [56]提出的特征函数来表示特征。对于文本分类，我们可以为每个词-类别组合定义一个特征函数如（3.1）所示：

(3.1)

其中表示 w 在文档 d 中出现的次数，表示文档 d 中的词数。

对于特征函数，记表示在训练集上关于的数学期望，表示关于模型与经验分布的数学期望，其计算公式如（3.2）～（3.3）所示：

(3.2)

(3.3)

如果模型能够获取训练数据中的信息，则就能假设这两个期望值相等，如公式（3.4）所示：

(3.4)

于是，若给定 k 个特征函数，它们分别为：，，…,，根据公式（3.4）就可以得到所求概率分布的 k 组约束。

因此，最大熵求解的问题就变成了满足一组约束条件的最优解问题，即：

(3.5)

(3.6)

求解这个最优解的经典方法是拉格朗日乘子算法，该算法证明了的取值符合下面的指数模型：

(3.7)

其中，是归一化因子，是模型的参数，可以看成特征函数的权值，它表征了对于模型的重要程度。上述式子使模型由求概率值转化为求参数值，一般的估计方法采用的是 Darroch 和 Ratcliff[57]的通用迭代算法，通过此算法得到具有最大熵分布的所有参数值，构造最大熵模型。

3.2 问句分类中的特征选取

中文问句特征和英文问句特征之间的不同表现在：中文问句的特征通常是以词语为基础的，词语又是由不同的字构成的，而字与字之间是没有任何标记的，因此，如何获取中文问句中的词成为提取特征的关键。所以，首先对中文问句进行分词和词性标记，然后把词和词性作为中文问句的最基本特征。另外，还有句法、词向量等特征，都是以分词和词性标记为基础提取出来的。本章提取了四种特征：词袋特征、词和词性特征、词向量特征、句法特征。

（1） 词袋特征

词袋顾名思义就是在不考虑词语的顺序、句法和语法等这些信息的情况下，仅仅将文本看成是一个词语的集合。由于问句都比较短，包含的词数比较少，因此在实际应用的过程中，本文提取问句中所有的词进行问句分类。例如，对于问句“*ACLU 的全称是什么*”，经过分词和词性标注后，提取词袋特征：

“*ACLU 的 全称 是 什么*”

因此，词袋特征是最基础的特征，同时也是最简单的特征，但是词袋特征会带来一定的噪音。

（2） 词和词性特征

通过观察训练数据发现，在分类中起作用的主要是实词，这是因为实词中通常含有较多有关类别的语法信息，而虚词（如介词、语气词、助词等）这些没有实际意义的词会带来一定的噪音，对分类造成干扰，因此，本文根据词性提取出问句中的实词和其对应的词性进行问句分类。例如：对于（1）中的例子进行词性标注后的结果为：“*ACLU/n 的/u 全称/n 是/v 什么/r*”，根据词性提取此问句中的实词及其词性特征为：

ACLU n 全称 n 是 v 什么 r

（3） 词向量特征

词向量是将词袋中的每个词用实数表示。传统最常用的词表示方法是One-hot Representation，这种方法把每个词表示成一个词表大小的向量，且该向量的绝大多数元素为 0，只有一个维度值为 1，这个维度就代表了这个词本身。但是采用这种方法来表示词，在解决某些任务的时候会出现维数灾难，且从这两个词的向量中看不出两者是否有关系。针对这些问题，一些研究者提出了分布式表示的词向量，它表示的是一种低维连续地实数向量。这种向量的提出解决了维度灾难和词汇鸿沟的问题。因此，被大多数研究者广泛采用。

（4） 句法特征

依存语法通过分析词语之间的依存关系揭示其句法结构，主张句子中核心动词是支配其他成分的中心成分，而其本身却不受其它任何成分的支配，所有受支配成分都以某种依存关系从属于支配者[58]。依存句法分析可以识别问句中的“定状补”、“主谓宾”等语法成分，并对各成分之间的依存关系进行分析。依存句法依据的标注体系及含义如表 3.1 所示。因此，本章利用问句中各成分间的依存关系抽取问句的句法特征。利用“主谓关系”、“核心”和“动宾关系”抽取问句中的主干，利用疑问词对应的依存关系抽取疑问意向词，得到一个五元组，其形式为<主语，核心词，宾语，疑问词，疑问意向词>，作为问句的句法特征。例如，对问句“什么部门负责对外贸易监控”进行依存分析后的结果如图 3.1 所示：

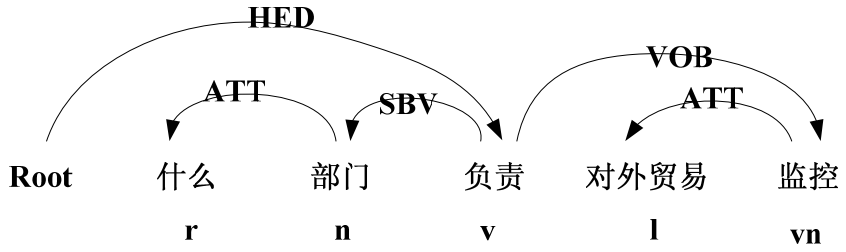

图 3.1 依存分析结果

如图 3.1 所示，Root 是支配整个句子的核心成分，“负责”为整个句子的核心。箭头所指的词是依存词，箭尾所指的词是核心词。“部门”依存于“负责”，其依存关系是“SBV”。“监控”依存于“负责”，其依存关系是“VOB”。因此，通过依存关系“SBV”、“HED”和“VOB”可以得到问句的主干“部

门负责监控”及其疑问词对应的依存关系“ATT”得到疑问意向词“部门”，于是，根据图 3.1 得到一个五元组<部门，负责，监控，什么，部门>。

表 3.1 依存句法标注体系及含义

关系	符号	关系	符号
定中关系	ATT	“的”字结构	DE
数量关系	QUN	“地”字结构	DI
并列关系	COO	“得”字结构	DEI
同位关系	APP	“把”字结构	BA
前附加关系	LAD	“被”字结构	BEI
后附加关系	RAD	状中结构	ADV
比拟关系	SIM	动宾关系	VOB
语态结构	MT	主谓关系	SBV
独立结构	IS	连动结构	VV
动补结构	CMP	关联结构	CNJ
介宾关系	POB	独立分句	IC
核心	HED	依存分句	DC
分析器无法确定的关系	NOT		

根据问句中各成分间的依存关系很容易提取出问句的主干，但是在抽取疑问意向词时，首先必须识别出问句中的疑问词。通过问句集的词性标注可以发现，疑问词的词性一般是“r”，但是有些代词如“它”，“其”的词性也是“r”，若只根据词性识别问句中的疑问词会导致抽取出的特征带有一定的噪音，因此为了更准确地抽取出问句中的疑问词，本文建立了一个疑问词表如表 3.2 所示，通过该表和词性来识别问句中的疑问词，最后，通过识别出的疑问词对应的依存关系抽取疑问意向词。例如：

“*什么/r 部门/n 负责/v 对外贸易/l 监控/vn*”

在此例中，首先根据疑问词表和词性“r”识别出疑问词“什么”，然后根据图 3.1 中“什么”对应的依存关系，得到疑问意向词“部门”。

表 3.2 疑问词表

多久	多少	何	何处	何时	几	哪	哪儿	哪个	哪里	哪些
如何	啥	什么	什 么	谁	为何	为 什	怎么	怎 么	怎 么	怎样

样　　么　　办　　样

抽取问句句法特征算法如 3.1 所示。设第 *i* 个问句经过依存分析的结果为 dependency，其中<,,,>表示第 *j* 个词的依存关系，表示第 *j* 个词，表示第 *j* 个词的词性，表示第 *j* 个词的依存关系，表示第 *j* 个词依存的词，*k* 表示第 *i* 个问句中包含的词数。设疑问词表为 $qword=\{qw_1,qw_2,\ldots,qw_m\}$，其中 qw_i 表示疑问词表中的第 *i* 个疑问词，*m* 表示疑问词表中包含的疑问词个数。则通过依存关系抽取问句特征得到一个五元组<*sbv,hed,vob,irg,irg_rel*>，其中 *sbv* 表示每个问句的主语，*hed* 表示每个问句的中心词，*vob* 表示每个问句的宾语，*irg* 表示疑问词，*irg_rel* 表示疑问意向词。

算法 3.1 ：抽取问句句法特征算法

输入：dependency={<,,,>,…, <,,,>},qword={,,…,}

输出：<sbv,hed,vob,irg,irg_rel>

过程：

```
While ：
    if ：
        hed;
    if 等于 hed：
加入到队列 openl 中;
end while
tmp;
while 队列 openl 不为空：
    if  tmp 的依存关系等于“SBV”：
        sbvw;
    else if  tmp 的依存关系等于“VOB”：
        vobw;
while ：
  if 等于“r”&&qword：
irg; irg_rel;
```

3.3 实验及结果分析

3.3.1 问句分类标准

问句的类型是由问句的分类标准决定的。不同的分类标准有不同的问句类

型。哈尔滨工业大学信息检索实验室在国外一些已有的英文问句分类标准的基础上，根据答案的类型和汉语自身的特点，定义了一个适合中文问句分类的标准，该标准得到了很多学者的认可，并且被国内研究中文问句分类的大多数学者广泛采用。本章采用该标准对问句进行粗粒度类别标注。表 3.3 给出了哈工大中文问句分类标准，包含 7 个大类，分别是：“描述”、“人物”、“地点”、“数字”、“时间”、“实体”、“未知”，每个大类根据实际情况又定义了一些小类，共 84 小类。

表 3.3 哈工大中文问句分类标准

大类	小类
描述（DES）	简写、表达、意思、方式、原因、定义、判断、其它
人物（HUM）	特定人物、机构团体、人物描述、其它
地点（LOC）	宇宙、城市、大陆、国家、省、河流、湖泊、山脉、大洋、岛屿、建筑、地址、其它
数字（NUM）	温度、面积、体积、重量、速度、频率、距离、钱数、数量、顺序、倍数、百分比、号码、时间长度、范围、其它
时间(TIME)	年、月、日、季节、时代、星期、节气、节假日、时间、时间范围、其它
实体（OBJ）	物质、动物、植物、微生物、身体、材料、机具、衣物、食物药品、货币、票据、语言、事件、疾病、艺术品、服务、文字作品、学术学科、计划规划、法律法规、职位头衔、职业行业、符号、奖励、刑法、类别、权利义务、颜色、宗教、运动娱乐、术语、其它

3.3.2 数据集

本章采用哈尔滨工业大学信息检索实验室问句分类数据集，该数据集共有 6296 个问句，其中训练集有 4981 个问句，测试集有 1315 个问句，数据集大类分布情况如表 3.4 所示：

表 3.4 训练语料和测试语料的问句分布情况

数据集	描述类	人物类	地点类	数字类	实体类	时间类	未知
训练集	786	333	936	1076	1242	598	10
测试集	153	178	390	244	194	153	3

3.3.3 实验过程

（1） 为了获得问句的词袋和词性特征，本章使用中科院分词系统[9]对中文问句分别进行分词和词性标注。

（2） 为了能把问句中独立的词的语义关系考虑进去，本章采用词向量表示词袋中的词。采用 Google 在 2013 年发布开源的工具 word2vec[10]来训练词向量，其中训练语料来自 60 年的人民日报，所训练词向量的维数为 200 维。然后，用训练好的词向量来表示问句中的词，得到每个问句的词向量特征。

（3） 本章采用哈尔滨工业大学信息检索实验室提供的语言技术平台共享包（LTP）[11]进行依存句法分析。然而，此平台不能使用外部字典，因此，为了以中科院的分词和词性标注为基础进行依存分析，本章在本地编译 ltp 源代码,编译后调用静态链接库 lib，通过 lib 来调用 LTP[12]对已经进行分词和词性标注的问句进行依存分析。然后根据依存分析的结果按照 3.1 的抽取算法获得问句的句法特征。

3.3.4 实验结果

基于最大熵模型的问句分类方法提取词袋特征、词和词性特征、句法特征、词向量特征进行问句粗分类实验，本实验采用准确率（P）[58]来评价粗粒度分类结果。四种特征对问句粗粒度分类准确率的影响如表 3.5 所示，四种特征对每个类别的影响如图 3.2 所示。

[9]https://github.com/NLPIR-team/NLPIR/tree/master/License
[10]https://code.google.com/p/word2vec
[11]https://www.ltp-cloud.com/demo/
[12]http://t.cn/zRcKbd4

表 3.5 四种特征对问句粗粒度分类准确率的影响

方法	准确率（%）
最大熵+词袋	88.59
最大熵+词+词性	87.59
最大熵+句法特征	88.67
最大熵+词向量	**88.75**

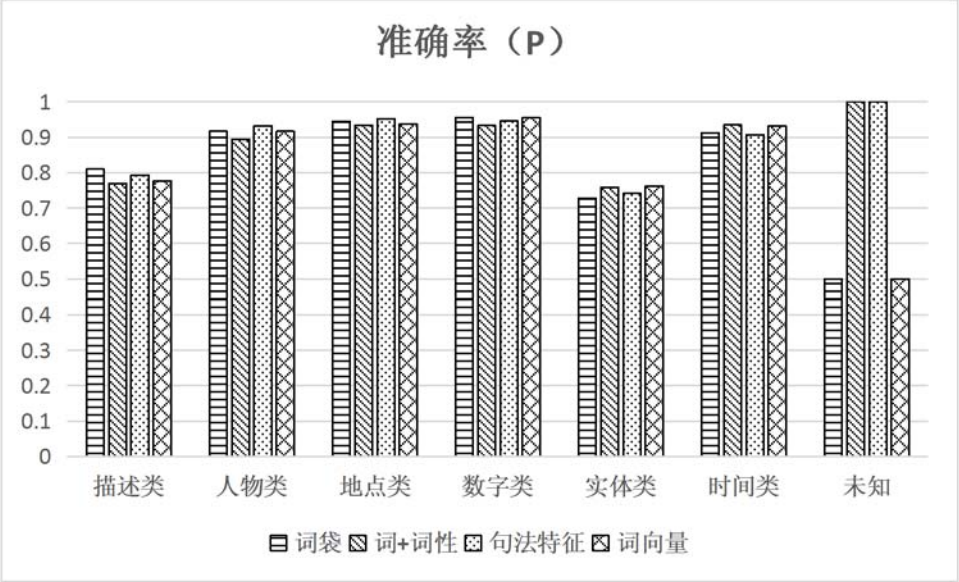

图 3.2 四种特征对每个类别的影响

由表 3.5 可以看出，相对其他三种特征而言，词向量特征对问句分类的性能较好。这是因为虽然词法信息和句法信息对问句分类性能的提升也有一定的优势，但是对于有些类型的问句来说，语义信息也是一个重要信息。而词向量特征是词语的分布式表示，它能反映词之间的相似关系且包含更多信息。另外，它还可以反映词的功能和上下文的语义信息，因此词向量对问句分类起着一定的作用。

由图 3.2 可以看出，词袋模型对描述类的分类性能较好，句法特征次之，词+词性特征对描述类的分类性能相比较而言较差。这是因为描述类的词法信息相对比较明显，因此基本的词袋模型就能使描述类达到较好的性能。句法特征与其他特征相比较而言，对人物类和地点类的区分能力较好。这是因为区分人物类和地点类的关键词大部分都是和疑问词相关的词，而这些词可以通过依存关系很好地提取出来，因此，句法特征能够提高人物类和地点类的分类性能。例如：

“武昌在哪个城市”

“什么部门负责对外贸易监控”

上述两个例句中“*哪个*”和“*什么*”是疑问词，通过依存分析的结果，提取与“哪个”有依存关系的词“*城市*”和与“*什么*”有依存关系的词“部门”，然后由这两个关键词可以判断出它们分别属于“地点类”和“人物类”。

词向量对数字类和实体类的分类性能较好；这是因为数字类和实体类的词法信息和句法信息不是很明显，它们大部分需要结合语义信息来判断最终的类别，而本章采用的词向量是是以分布式的方法来表示词语，它能够包含更多信息且可以反映词的功能和上下文语义信息，因此它对区分数字类和实体类具有较好的优势。词性特征对时间类的区分能力相对其他特征而言较好一些。这是因为时间类的词性特征特别明显。例如：

“人类 n 有史以来 l 第一 m 次 qv 登 v 月 n 是 v 哪 r 年 qt”

上述例句中“*年*”的词性是“*qt*”，而“*qt*”表示时间量词，因此可以根据词性判断该问句属于时间类，因此词性对于提高时间类的性能具有一定的优势。

3.4 本章总结

本章介绍了基于最大熵模型的问句粗粒度分类方法，且介绍了最大熵模型应用到问句分类中的原理，然后，综合考虑不同类别下的问句特点，分别提取四种特征：词袋、词和词性、句法、词向量，最后分别利用最大熵模型在哈工大问句集上进行粗粒度分类实验，实验结果表明，与其他三种特征相比，词向量特征在问句粗分类上表现出较好的性能，准确率达到 88.75%。

4 基于 Bi-LSTM 的问句分类

近年来，深度学习在 NLP 领域中得到了广泛应用，一些研究人员利用深度学习的方法对问句分类进行了探索，并取得了一定的成就。本章提出了一种基于 Bi-LSTM 的问句分类模型，该模型融合词、词性和词的位置特征生成词语的嵌入表示，利用 Bi-LSTM 自动学习问句的语义表示，采用 Adam 算法进行梯度更新，该算法能很快收敛，并快速找到参数更新中正确的目标方向，最大程度地最小化损失函数。最后利用 Softmax 对问句进行分类。本章的方法能够在输入和输出序列之间的映射过程中充分利用上下文信息，有利于全面理解问句的语义信息。

4.1 LSTM 模型

循环神经网络（RNN）已被广泛用于处理可变长序列输入，长距离历史信息存储在一个递归的隐藏向量，它依赖于前一个隐藏向量。长短期记忆模型（LSTM）是一种循环神经网络（RNN），它能够记住很长时期内的信息，从而避免 RNN 无法解决的长期依赖问题。LSTM 关键的部分是记忆单元，它由输入门，遗忘门和输出门，来实现保护和控制信息。输入门控制加入记忆单元的信息的数量，遗忘门控制通过记忆单元的信息，输出门控制输出信息。

LSTM 神经元结构如图 4.1 所示，这三个门在 LSTM 传播过程中的实现公式如（4.8）～（4.13）所示：

(4.8)
(4.9)
(4.10)
(4.11)
(4.12)
(4.13)

其中是激活函数，表示向量之间的点乘运算，表示时间步。输入门、遗忘门和输出门是由前一个状态和当前输入共同决定的，所提取的特征作为候选存储单元。由候选存储单元和前一个存储单元分别乘以各自的权重输入门和遗忘门，再相加得到当前存储单元。最后，由输出门和当前存储单元计算得到 LSTM 单元的输出。

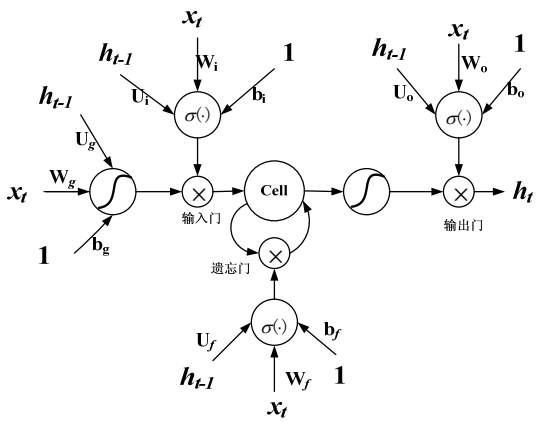

图 4.1 LSTM 神经元结构

4.2 Bi-LSTM 模型

单向 LSTM 模型虽然能够解决 RNN 中的梯度消失问题，但是它只能捕捉前词的特征。Bi-LSTM 模型通过在两个方向上处理序列来捕捉前词信息和未来的上下文信息，并生成两个独立的 LSTM 输出向量序列，一个正向处理输入序列，另一个处理反向输入。每个时间步的输出是来自两个方向的两个输出向量相加得到的。即。Bi-LSTM 模型的结构如图 4.2 所示：

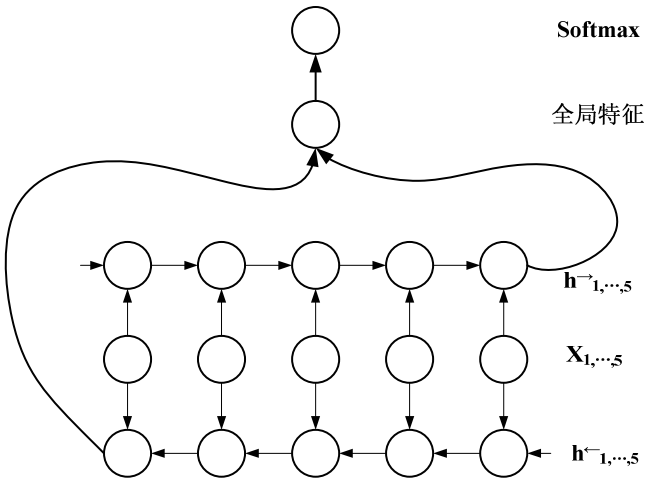

图 4.2 Bi-LSTM 模型结构图

4.3 特征融合

在基于 Bi-LSTM 的问句分类模型中，本章提取词、词性和词语的位置信息来提高问句分类的性能。首先将这三种特征进行融合，生成词语的嵌入表示。融合方法有：特征拼接、特征相加和特征乘积。设第 i 个问句为其中，为问句中第 i 个词，表示第 i 个问句的长度。词语对应的词向量，对应的词性向量，对应的位置向量为。则特征拼接是将、、这三种特征向量直接拼接生成第 i 个词语的嵌入表示，三者的维度满足如下条件：.。其拼接公式如（4.14）所示；特征相加是将、、这三种特征向量对应维度的值相加得到第 i 个词语的嵌入表示的每一维度上的值，但是进行特征相加时，必须满足三者的维度相同，即 $l=m=n$。其计算公式如（4.15）所示；特征乘积是将、、对应维度上的值相乘得到第 i 个词语的嵌入表示的每一维度上的值，且三者的维度满足关系：$l=m=n$。其计算公式如（4.16）所示。由于特征拼接简单实用，因此本章选择的是特征拼接的方式进行特征融合。

(4.14)

(4.15)

(4.16)

4.4 基于 Bi-LSTM 的问句分类模型

基于 Bi-LSTM 的问句分类模型如图 4.3 所示，该模型主要由语料预处理、词嵌入和分类三个模块组成。语料预处理模块主要是完成对生语料的处理，包括分词、词性标注等操作。设表示第 i 个问句，则经过预处理后，生成词语序列，词性序列，位置序列，其中 k 表示第 i 个问句中词的个数，表示第 i 个问句中的第 j 个词，表示第 j 个词的词性，表示第 j 个词语的位置。词嵌入模块主要完成特征向量化和特征向量拼接。首先将预处理模块生成的序列中的、、分别向量化，然后将得到的词向量、词性向量和词语位置向量进行拼接生成词语的嵌入表示。分类模块主要由 Bi-LSTM 隐藏层、最大池化层（Max Pooling）和 Softmax 层三部分组成。Bi-LSTM 隐藏层生成问句的分布式表示，通过 3 个门函数对问句的状态特征进行正向和反向计算，分别得到正向输出和反向输出，然后将和进行求和计算得到第 j 个词语的嵌入表示对应的隐藏层的最终输出。接下来，对

第 *i* 个问句中的所有词嵌入对应的隐藏层的输出进行 Max Pooling，生成第 *i* 个问句的全局表示，最后，在 Softmax 层进行问句分类，输出每个问句属于某一类别的概率。

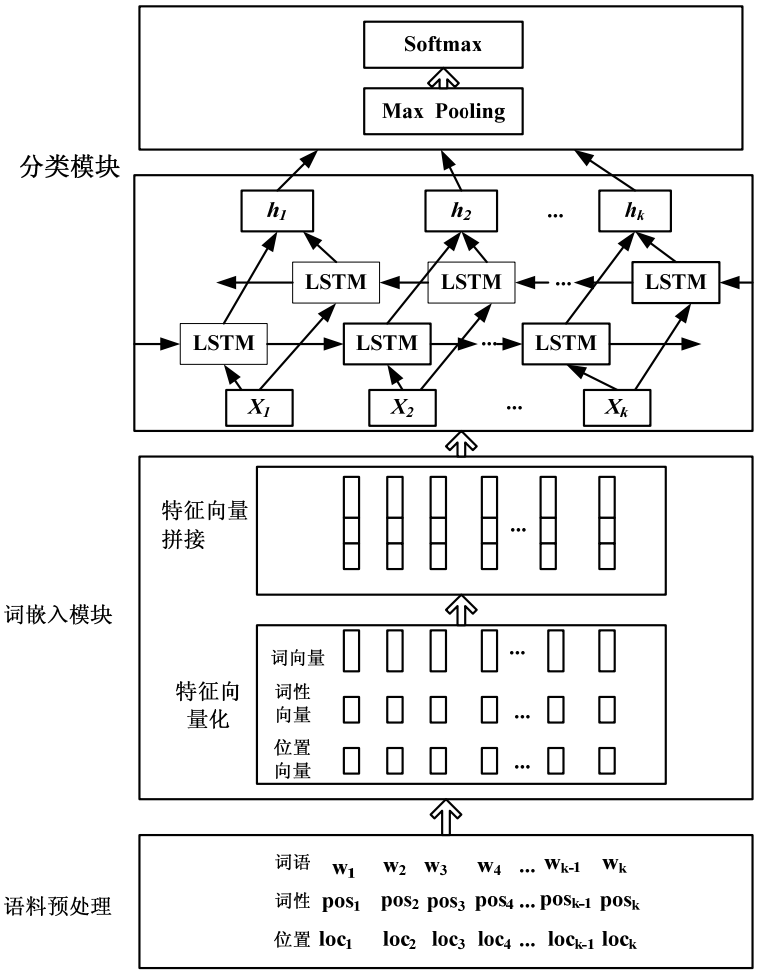

图 4.3 基于 Bi-LSTM 问句分类模型

4.4.1 语料预处理模块

在分类模型中，数据的质量直接影响着分类的性能。词语作为问句的基本组成单元，存储了问句的基本特征，此外，词性信息和词语的位置信息也对判断问句类型有着一定的指导意义。在预处理模块，本章考虑句首、句中和句尾三种位置信息，分别用“1，2，3”表示。语料预处理的算法如 4.1 所示：

算法 4.1：语料预处理算法

输入：问句集，其中 N 表示问句集中的问句数，问句集中的第 i 个问句

输出：、、

过程：　while ：

1. 将进行分词和词性标注，得到的词语序列;
2. 抽取的第 j 个词的词性，得到的词性序列
3. 根据的第 j 个词在问句中的位置，分别用数字“1”、“2”、“3”表示，得到的词语位置序列;
4. i++;
5. end while

4.4.2 词嵌入模块

词嵌入作为词语的一种稠密低维的连续表示，能够有效地表示一个词语的语义句法等信息[37]。词嵌入模块主要由特征向量化和特征向量拼接两部分组成。词嵌入生成的算法如 4.2 所示：

算法 4.2：词嵌入生成算法

输入：第 i 个问句的词语序列，词性序列，词语位置序列，其中 k 表示问句的长度

输出：（）

过程：对第 i 个问句中的词语序列、词性序列和词语位置序列中的、、：

1. while ：
2. 将、、向量化，得到的词向量，的词性向量，的位置向量;
3. 将、、依据公式（4.14）进行拼接生成第 j 个词语的嵌入表示;
4. end while

4.4.3 分类模块

分类模块主要由 Bi-LSTM 隐藏层、Max Pooling 层和 Softmax 层三部分组成。首先将词嵌入生成模块得到的作为 Bi-LSTM 隐藏层的输入，通过 3 个门函数分别对进行正向和反向计算。设正向计算得到的输出和反向计算得到的输出，则将和求和得到对应的隐藏层的最终输出，其计算公式如（4.17）所示：

(4.17)

对于第 *i* 个问句，通过 Bi-LSTM 隐藏层得到一个输出值序列，其中表示对应的 Bi-LSTM 隐藏层输出，*k* 表示第 *i* 个问句的长度。将作为 Max Pooling 层的

输入，生成第 *i* 个问句的全局表示，其计算公式如（4.18）所示：

(4.18)

最后通过 Softmax 层对问句进行分类。在 Softmax 层依据假设函数估算问句属于每个类别 *t* 的概率公式如（4.19）所示：

(4.19)

其中 *c* 表示类别数。

假设函数的计算公式如（4.20）所示：

(4.20)

其中，模型参数。由于共有 c 个类别，因此 Softmax 层最终会输出一个 c 维的向量，且向量每一维的值就是问句属于各类别的概率值。分类算法如 4.3 所示：

算法 4.3：分类算法

输入：第 i 个问句的词嵌入序列：

输出：

过程：沿网络前向传播：

当时：

1. 对每个词嵌入：

   Bi-LSTM 隐藏层依据公式（4.8）～（4.13）分别对进行正向和反向计算，得到正向输出和反向输出;

将和依据公式（4.17）求和，得到对应的 Bi-LSTM 隐藏层的输出;

结束循环

2. 第 i 个问句的输出序列为;
3. 将作为 Max Pooling 层的输入，依据公式（4.18）生成问句的表示;
4. 将作为 Softmax 层的输入依据公式（4.19）和（4.20）求得一个输出向量，且该向量的每一维就是第 i 个问句在该类别的概率值;

代价函数：（其中 z 表示真实值）;

反向传播更新权值、b：

;

（其中，）

4.5 实验及结果分析

4.5.1 数据集

本章采用第三章的问句分类数据集和问句分类标准进行实验。

4.5.2 Bi-LSTM 模型参数设置

本章采用固定问句长度，且长度设置为最长问句长度，不足用“0”填充。随机初始化词向量、词性向量和词语位置向量，且在训练过程中这些向量是不变的，其维度分别为：150 维、100 维、100 维。Bi-LSTM 模型隐藏层有 100 个 LSTM 单元组成，采用学习率自适应算法（Adam）对参数进行梯度更新。训练时采用批量训练的的方法，且批量大小取 128。网络训练的最大训练轮数设为 2000 轮，每轮训练中，dropout 方法的 dropout rate 取 0.4。

4.5.3 实验结果

基于 Bi-LSTM 的问句分类融合词、词性和词语位置特征生成词语的嵌入表示，通过 Bi-LSTM 模型进行问句粗粒度分类实验。实验采用 3 种评价标准：准确率（P）、召回率（R）、F1 值（F1）[58]评价词语位置特征对每个类别的影响，如表 4.1 所示。为了方便，本文将“Bi-LSTM+词语+词性”用 model1 表示，“Bi-LSTM+词语+词性+词语位置”用 model2 表示。另外，本章将基于 Bi-LSTM 的问句粗粒度分类结果与基于最大熵模型的问句粗粒度分类结果以及国内其他

中文问句粗粒度分类结果进行了对比，如表 4.2 所示。

表 4.1 词语位置特征对每个类别的影响

类别	model1			model2		
	P（%）	R（%）	F1（%）	P（%）	R（%）	F1（%）
描述类	83.87	84.97	84.42	86.36	86.93	86.64
人物类	**84.04**	**88.76**	**86.33**	**93.49**	**88.76**	**91.06**
地点类	95.80	87.69	91.57	96.24	91.79	93.96
数字类	96.57	97.54	97.14	96.75	97.54	97.14
实体类	**75.23**	**84.54**	**79.61**	**81.08**	**92.78**	**86.54**
时间类	94.04	92.81	93.42	95.97	93.46	94.70
未知	0.00	0.00	0.00	33.33	33.33	33.27

由表 4.1 得出，融合词语位置特征前，采用 Bi-LSTM 模型对问句进行粗粒度分类时，微平均准确率达到 90.27%；融合词语位置特征后，对问句进行粗粒度分类时，微平均准确率达 92.38%。这说明将词语位置特征添加到词嵌入生成过程中对分类性能有一定的提升。而且由表 4.1 可以看出，词语位置特征对人物类和实体类有明显的提升，准确率分别提升了 9.5 个百分点和 5.9 个百分点。这是因为决定问句属于人物类或者实体类的关键词的位置较明显。例如：

“*HP 是 哪个 公司 的 简称*”

上述例句的关键词“公司”和“简称”的词性相同，位置不同。而关键词“*简称*”又是描述类的特征之一，若不考虑关键词的位置，不容易判断该问句的类别，因此关键词的位置起到了一定的作用，再结合 Bi-LSTM 自身固有的优势，就能正确判断该句的类别。由表 4.1 可知，数字类提升的效果不明显，这是因为数字类的问句中会含有“面积”、“邮编”、“号码”、“区号”等特征明显的实体。因此，添加位置信息对数字类问句提升效果不明显。

表 4.2 中文问句粗粒度分类工作对比

	特征	模型	数据规模	类别数目	数据集	P(%)
文勖等，2006	主干+疑问词+附属成分	Bayes	9600	7	文献[23]+中科院问句集	86.62

田卫东等，2010	自学习规则	Bayes	4280	6	哈工大问句集	84.00
张宇等，2005	词+词性+TF-IDF	Bayes	4280	7	哈工大问句集	72.40
余正涛等，2005	词+词性+语块句法特征+隐含语义特征	SVM	1500	6	收集的汉语问句	88.70
李超等，2016	自我学习特征	LSTM+CNN	9600	6	哈工大问句集+NLPCC2015+复旦大学问句集	93.08
孙景等，2007	疑问词+句法结构+疑问意向词+首义原	ME	5613	7	哈工大和中科院问句集	92.18
EM-Embedding	**词向量**	**ME**	**6296**	**7**	**哈工大问句集**	**88.75**
Bi-LSTM	**词+词性+位置**	**Bi-LSTM**	**6296**	**7**	**哈工大问句集**	**92.38**

由表 4.2 可以看出，在同一数据集上进行粗粒度分类实验，基于 Bi-LSTM 的问句分类方法高于最大熵模型的分类方法，且提升了 3.6 个百分点。这是因为传统方法对问句分类大都采用人工制定提取特征的策略，灵活性不高，而且还有一定的局限性。虽然词向量特征包含更多的语义信息，但是它会出现词语歧义的问题，因此会影响问句分类的性能。而 Bi-LSTM 不仅能自主学习到当前词的完整的过去，还能学习到它的未来上下文信息，这样提取的问句特征能够全面理解问句的语义信息，提升问句分类的性能。

通过表 4.2 发现，本章提出的问句分类模型尚不能达到目前最好分类结果。可能的原因为：本章实验中使用的数据量较少，而 Bi-LSTM 模型中的参数数目较大，在实验数据较少时训练出较好的分类器会有一定的难度。而李超等[36]结合单向 LSTM 和 CNN 两种学习模型的优点，使用较多的实验数据自主学习问句的深层句法语义特征，进一步挖掘出问句的更多深层语义信息，因此，能够训练出性能较好的分类器。

4.6 本章总结

虽然传统方法对问句粗粒度分类性能有一定的提升作用，但是它们需要人工提取特征，既费时又耗力。因此，为了弥补传统方法带来的不足，本章提出了一种基于 Bi-LSTM 的问句分类方法，该方法根据自身的优势能够全面理解问句的语义信息，对提升问句的性能有一定的帮助。首先，该方法融合词、词性和词语位置特征生成词语的嵌入表示，通过 Bi-LSTM 隐藏层提取问句的分布式表示，最后通过最大池化层和 Softmax 层分别提取问句全局特征和对问句进行粗粒度分类。在哈工大数据集上进行实验，粗粒度分类准确率达到 92.38%，比传统方法提升了 3.6 个百分点。

5 问句分类在知识库问答系统中的应用

本章主要研究利用知识库来回答简单问题，所谓的简单问题是指由实体和一个二元关系描述组成的问句。面向知识库问答系统与基于文档的问答系统不同，它的答案是以结构化三元组的形式存在的，即实体1-关系-实体2。为了方便区分，本章节将实体1称为“主语”，关系称为“谓语”，实体2称为“宾语”。将问句中与相应的知识库查询三元组中的实体1对应的子串称为问句的话题短语。例如“*你知道知母是什么纲的吗？*”，该问句对应的知识库查询三元组为“*知母 /// 纲 /// 单子叶植物纲 Liliopsida*”，其中知识库实体“*知母*”就是问句中的一个子串，即问句的话题短语。

基于知识库的问答系统的实现步骤如图5.1所示。本章分三部分来实现知识库问答系统：问句分析、答案抽取和答案排序。其中问句分析包括问句话题短语检测、问句谓词识别和问句分类三部分；答案排序通过特征提取和排序两部分组成。问句话题短语检测是识别出包含在知识库中的问句话题短语，问句谓词识别是在识别出问句话题短语后，根据依存关系和规则识别出问句中的谓词，将问句结构化为“话题短语实体-谓词”的形式，并将其应用到答案排序的特征提取阶段。而问句分类则被作为其中的一种特征应用到答案排序中。

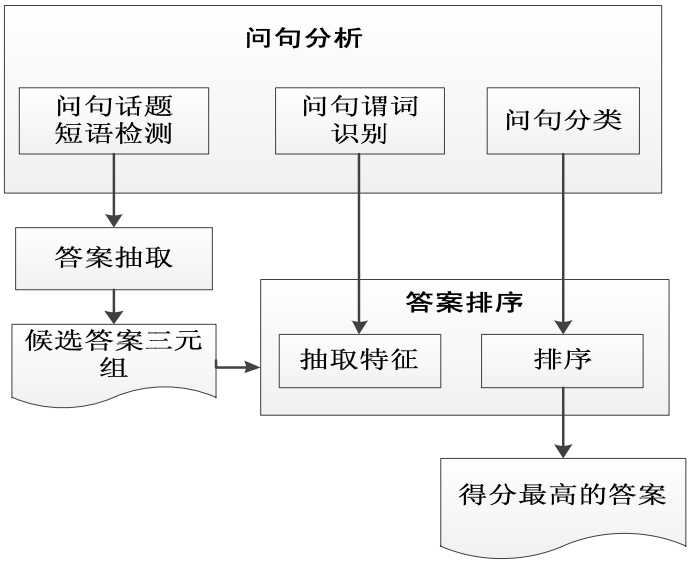

图5.1 基于知识库问答系统实现步骤图

5.1 问句话题短语检测

问句话题短语检测分两步来完成。第一步，首先使用实体短语词典检测出现在问句中的所有短语。以词典中的每个短语为子串，问句为目标串，采用模式匹配的方法检测出问句中的所有短语。第二步，检测话题短语。首先，根据问句-答案对标注一个训练集，即是对第一步中识别出的所有短语进行“0”和“1”的标注。训练集的形式为<phrase,label>。其中 phrase 表示第一步中识别出的短语，label 表示标签“0”或“1”。从标注好的训练集中提取特征，利用随机森林算法检测出每个问句中的话题短语。

5.1.1 基于模式匹配的问句短语检测

模式匹配包括朴素的模式匹配和 KMP 匹配两种算法[59]。本章采用朴素的模式匹配算法。首先，本章从知识库中抽取所有的实体短语，将抽取出的实体短语去重，得到一个实体短语词典。然后，将短语词典中的每个短语作为模式串，而问句作为目标串，来检测问句中出现的所有短语。

设问句集，其中表示第 *i* 个问句，N 表示问句集中的问句数。短语实体集 P={}其中表示第 *j* 个短语，m 表示短语实体集中的短语数，则采用朴素模式匹配查找问句中的所有短语算法如 5.1 所示：

算法 5.1：朴素模式匹配算法

输入：

输出：f,其中 f 是一个布尔值，当模式匹配到的短语是知识库词典中的短语时，该值为 true

过程：1.while ：

2. while ：

3. , f=false;

4. while a&& k：

5. if ==： a++;b++;

6. else：a=0;b=k+1;k++;

8. if a== f=true;

9. else： f=false;

10. j++;

11 . i++;

5.1.2 基于随机森林的话题短语检测

本章将识别问句中的话题短语看成一个二分类问题，类别标签为“0”和“1”，其中“1”表示是话题短语，“0”表示不是话题短语。问句话题短语检测分成两步来完成：第一步，构建分类训练集；第二步，从构建好的分类训练集中提取特征，训练一个随机森林分类模型，根据树分类器投票数得到每个短语的类别。

（1） 构建分类训练集

分类训练集是根据问答系统训练集中的答案（也是知识库三元组中的宾语）构建的。设知识库三元组，其中分别表示第 c 个三元组的主语，谓语，宾语；；第 i 个问句中的所有短语，其中表示第 i 个问句中的第 j 个短语，k 表示第 i 个问句包含的短语数，表示第 i 个问句对应的答案。则构建训练分类集（表示第 j 个短语的标签）算法如 5.2 所示：

算法 5.2：构建分类训练集算法

输入：；

输出：

过程：
```
1.  while ：
2.      if 中的等于：
3.          while ：
4.              if s：将
5.              else：将;
6.               j++;
7.  else：a++;
```

（2） 特征提取

从构建好的训练分类集中提取当前短语前词和后词特征以及计算当前短语在问句语料库中的逆文档频率和在知识库三元组的谓语语料库中的逆文档频率、其他特征。

1）当前短语前词特征

所谓的前词是指当前短语前面的短语。首先从构建好的所有训练样例中生成一个前词-概率词典。如果当前短语在句首，则用一个特殊的字符串“HEAD”来表示当前短语的前词特征。我们抽取所有短语的前词，统计前词在训练样例

中的概率，从而得到一个前词-概率词典。用表示前词-概率词典的第 *i* 个元素，是第 *i* 个前词。则的定义如公式（5.1）所示：

(5.1)

其中，表示前词的概率值，表示每个训练样例中前词是的正样例数，表示每个训练样例中前词是的负样例数。

另外，由于前词-概率词典是根据训练样例构建的，而测试样例中的短语有可能在训练样例中没有出现过，因此对于这些短语本章节进行了 Good-Turning 的平滑处理来求其概率。其公式如（5.2）所示：

(5.2)

其中表示发生次的概率，表示发生 *r+1* 次的前词数，表示发生 r 次的前词数。

2）后词特征

所谓的后词就是当前短语后面的邻接短语。除了将前词换成后词外，后词特征的提取和前词特征的提取方法是相同的。如果当前短语在句尾，就用特殊字符串“END”作为当前短语的后词特征。

3）计算短语在问句语料库中的逆文档频率

设训练问句语料库为，表示第 *i* 个问句，*n* 表示问句语料库中包含的问句数。则逆文档频率的计算公式如（5.3）所示：

(5.3)

其中，表示包含第 *j* 个短语的问句数

4）计算短语在谓语语料库中的逆文档频率

从知识库的三元组中抽取所有的谓语组成谓语语料库，m 表示语料库中包含的谓语数，表示第 *i* 然后依据公式（5.4）计算每个短语在谓语语料库中的逆文档频率。

(5.4)

其中，表示包含第 *j* 个短语的谓语数。

5）其他特征

从构建的分类训练集中观察到，短语的长度及其在问句中的位置也起着重要的作用，因此，本节还提取了每个短语的长度及其位置作为检测问句中的话题短语的特征。

5.2 候选答案抽取

经过问句话题短语检测后，得到一个话题短语词典。根据每个问句的话题短语，从知识库中抽取主语是该话题短语的所有三元组，并将其作为问句的候选答案集合。

开放域知识库涉及的领域比较广泛，收录的数据比较齐全，诸如人物名称、书名等这些类型的实体在知识库中虽然同名但代表不同领域，因此为了将它们区分开，在把这些数据收录到知识库中时，通常会在这些实体的后面加上一些描述信息。而用户提出的问题一般都比较简略，后面不会出现这些描述信息，因此考虑到答案的完备性，无论知识库三元组的主语是否有描述信息，只要该主语与问句的话题短语匹配，本章均将其作为问句的候选答案。例如：“王伟是什么职业呀？”问句中并没有描述有关王伟的任何信息，而知识库中却按不同身份收录了“王伟”的有关信息，因此，将所有主语是“王伟”的三元组均作为该问句的候选答案集。该问句的候选答案集合如图 5.2 所示。

王伟(山东农业大学讲师) ||| 主要成就 ||| 法学硕士学位
王伟(山东农业大学讲师) ||| 性别 ||| 女
王伟(山东农业大学讲师) ||| 专业 ||| 刑法学
王伟(山东农业大学讲师) ||| 籍贯 ||| 山东省泰安市
王伟(山东农业大学讲师) ||| 国籍 ||| 汉族
王伟(山东农业大学讲师) ||| 出生年月 ||| 1982年2月
王伟(山东农业大学讲师) ||| 职业 ||| 教师
王伟(濮阳河务局副局长、党组成员) ||| 别名 ||| 王伟
王伟(濮阳河务局副局长、党组成员) ||| 中文名 ||| 王伟
王伟(濮阳河务局副局长、党组成员) ||| 民族 ||| 汉族
王伟(濮阳河务局副局长、党组成员) ||| 出生地 ||| 河南省濮阳县
王伟(濮阳河务局副局长、党组成员) ||| 出生日期 ||| 1957年9月
王伟(八十九团党委常委、武装部长) ||| 别名 ||| 王伟
王伟(八十九团党委常委、武装部长) ||| 中文名 ||| 王伟
王伟(八十九团党委常委、武装部长) ||| 民族 ||| 中国
王伟(八十九团党委常委、武装部长) ||| 出生地 ||| 汉族
王伟(八十九团党委常委、武装部长) ||| 职业 ||| 海南政法职业学院教授

图 5.2 问句“王伟是什么职业呀？”的候选答案集

设知识库三元组，其中分别表示第 c 个三元组的主语、谓语、宾语，M 是知识库中包含的三元组数，第 *i* 个问句的话题短语。则抽取知识库中与问句话题短语匹配的候选答案三元组 cad 的算法如 5.3 所示：

算法 5.3 ：抽取候选答案三元组算法

输入：；

输出：cad

过程：1. while ：

2. 截取第 j 个三元组的主语的描述信息前面的字符串 S（如：王伟（山东大学讲师

```
王伟）
3.          if   S==：
4.
5.          if 包含“《”：
6. 截取书名号里面的字符串
7.              if   S==：
8.
9.          j++;
```

5.3 答案排序

答案排序是问答系统中的最后一步，也是匹配正确答案的关键所在。而问句结构化成为在知识库中匹配正确答案的基础。因此，问句话题短语识别出来后，将问句中的话题短语用一个特殊的字符串“entity”来代替，然后采用第3章中的哈工大LTP平台对问句进行依存分析，提取出与“entity”有依存关系的词、中心词、疑问词和疑问意向词作为问句的谓词，将问句映射为“话题实体-谓词”的结构化形式。最后，基于问句谓词和抽取出的候选三元组中的谓语提取特征，对候选答案集进行排序。答案排序的过程如图5.3所示：

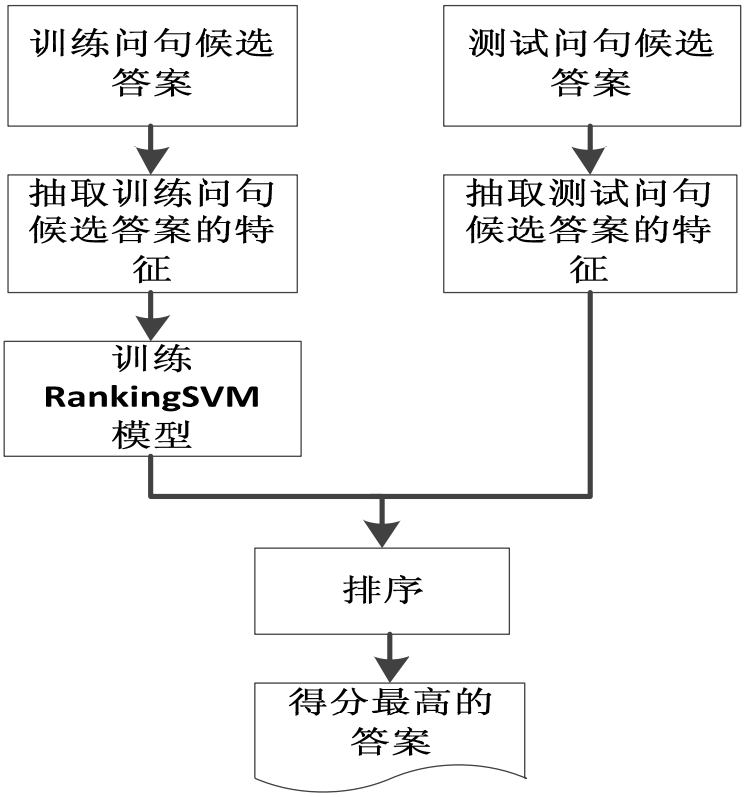

图 5.3 答案排序具体过程

另外，使用词向量计算语义相似度时，会出现类似“什么时候”和“日期”、“在哪儿”和“地点”之间的相似度会比预期的要低得多的问题。因此，本章将指定问句类型规则的词直接加入到问句谓词中。其规则如表 5.1 所示：

表 5.1 采用的规则

问句谓词中的词	要添加的词
什么时候，何时，多久	时间，日期
在哪，在哪里	地点，位置
多少钱	价格
多少页	页数
几岁	年龄

5.3.1 基于问句谓词和候选三元组的谓语的特征提取

（1） 相似度特征

本章采用词向量的余弦相似度来计算问句谓词和候选三元组的谓语的相似

度。在计算时，使用问句谓词的局部子序列与候选三元组谓语的词序列求相似度。设问句谓词的词语序列为，其中表示问句谓词中的第 *i* 个词语，*n* 表示问句谓词中包含的词语数量。候选三元组的谓语的词序列为，其中表示候选三元组谓语中的第 *i* 个词语，*l* 表示中词语的数量。则计算 p 和 q 间的相似度特征的算法如 5.5 所示：

算法 5.5：计算相似度算法

输入：;

输出：x，其中 x 表示问句谓词和候选三元组的相似度值

过程：;

```
1. while ：
2. tmp0.0,max0.0;
3.       while ：
4.             if 等于：tmp1.0;
5.             else：tmp 和的余弦相似度值;
6.             if : max;
7.           j++;
8.       xx+max;
9.       i++;
```

（2） 编辑距离特征

基于编辑距离的这个特点考虑，本章将问句谓词采用替换、插入或删除三种操作转成候选三元组的谓语，并记录下转化过程中的最小操作次数，即编辑距离。编辑距离越大，说明问句谓词和此候选三元组的谓语差异越大，成为正确答案的可能性就越小。反之，该候选三元组越有可能成为正确答案。例如问句谓词中的“上市时间”和候选三元组中的谓语“上市日期”，将问句谓词中的“时”替换为“日”，“间”替换为“期”后，问句谓词就转化成了候选三元组的谓语，且进行了两次替换操作，所以编辑距离为 2。计算问句谓词 S_1 和候选三元组的谓语 S_2 的编辑距离算法如 5.6 所示：

算法 5.6：编辑距离算法

输入： S_1,S_2

输出：d[n][m]

过程： n ←的长度, m ←的长度;

```
1.  if  n==0
2.      输出 m;
3.  if  m==0
4.      输出 n;
5.  while ：
6.      d[i][0]←i;
7.      i++;
8.  while ：
9.      d[0][j] ←j;
10.     j++;
11. while ：
12.     ←获取的第(i-1)个位置字符;
13.     while ：
14.         ←获取的第(j-1)个位置字符;
15.         if ==： tmp←0;
16.         else： tmp←1;
17.         d[i][j]←min{d[i-1][j]+1, d[i][j-1]+1, d[i-1][j-1]+tmp};
18.     j++;
19.   i++;
```

（3） 词共现特征

通过观察训练数据中的问句和正确答案发现，问句谓词中的一些词语与正确答案对应的三元组的谓语中的词语共现的次数较高。在训练数据中，本章统计了问句谓词中的每个词语和正确答案的谓语的每个词语共现的次数，得到一个词共现表。其中，表示第 *i* 个共现词语，表示其对应的共现次数,*k* 为共现词典的大小。这说明了当问句谓词中出现这些词语时，若候选三元组的谓语中出现了与其共现频率较高的词语，则该三元组成为正确答案的可能性就越高。计

算问句谓词 q 和候选三元组的谓语 p 共现频率 *c* 算法如 5.6 所示：

算法 5.6 ：共现频率算法

输入：;;

输出：c

过程： count←0

```
1.  while（n 表示 q 中包含的词语数）：
2.      while （l 表示 p 中包含的词语数）：
3.          if 不等于：
4.  S←将和连接;
5.              if d 的共现词集合中包含 S：
6.                  count←count+;
7.              else：
8.                  count←count+;
9.          j++;
10.     i++;
11. ;
```

（4） 分类特征

问句的类型指定了答案的类型，当候选答案三元组的类型和问句的类型一致时，该三元组成为正确答案的可能性就越大。该特征提取分三步来完成：首先，根据第四章的 Bi-LSTM 分类模型获取问答评测方提供的一万条训练问句的类别，由于问句类别即是答案类别，从而获得一万条问句对应的答案的类别。然后，以已获取类别的答案为训练集，提取词、词性和位置特征生成词嵌入，利用第四章的 Bi-LSTM 模型对第 2 节抽取的候选答案三元组的宾语进行分类，在 softmax 层获得候选答案三元组的宾语属于每个类别的概率。最后，根据问句的类别和候选答案三元组的宾语属于每个类别的概率，得到每个问句对应的候选答案三元组在该问句类别下的概率值。

设第 i 个问句对应的类别为，第 i 个问句对应的第 *j* 个候选答案三元组的宾语属于每个类别的概率为，其中表示第 k 个类别，表示宾语属于第 *k* 个类别的概率，*k* 表示类别数。提取分类特征的算法如 5.7 所示：

算法 5.7：分类特征提取算法

输入：;

输出：probality，其中 probality 表示候选答案三元组的宾语在其对应问句类别下的概率值

过程：对于第 j 个候选答案三元组：

```
1.  while ：
2.      if ==：
3.          probality;
4.      else：m++;
```

5.3.2 基于 Ranking SVM 的答案排序

（1） Ranking SVM

Ranking SVM[60]是一个学习排序函数，它在信息检索中得到了广泛的应用。该排序函数的输出是一个数据样例的得分，从中得到数据样例的局部排序。即目标函数输出一个得分，使得,对于任何。评价排序函数 F 通常通过它的排序接近的程度，为数据的最佳排序，R 被认为是严格的排序，它代表对数据集 D 中的所有和，存在两种情况：或者。

局部排序函数 F 利用 SVM 的技术从排序 R 中学习得到的。设线性排序函数 F 满足条件如公式（5.5）所示：

(5.5)

其中权重向量 w 是根据学习算法更新的。如果对所有的，存在一个函数 F（由权重向量表示）满足公式（5.5），则我们就称这个排序 R 是线性排序的。

设训练数据形如，其中每个样本包括两个特征向量和一个标签排在前面。则 Ranking SVM 可以被公式化为如下 QP 问题，如公式（5.6）～（5.8）所示：

minimize： (5.6)

subject to: (5.7)

i=1,…,m (5.8)

其中和是特征向量对中的第一个和第二个特征向量，表示 L2 范数，m 表示训练样本的数量，是一个系数。

（2） 答案排序

本章经过问句话题短语检测和候选答案抽取后，根据检测出的问题话题短语在知识库中抽取了问答系统训练问句的候选三元组和待解答的测试问句的候选三元组。

利用训练问句的标准答案，对抽取的训练问句的候选三元组进行标签标注。

这些标签分别为“0”表示是该问句的标准答案，“1”表示不是该问句的标准答案。然后，提取训练问句的候选三元组的特征，训练 Ranking SVM 模型。最后，提取问答系统的测试问句的候选三元组的特征，利用训练出的 Ranking SVM 模型对测试候选三元组进行排序，将排列在前面的候选三元组的宾语作为正确答案返回。训练 Ranking SVM 模型的数据格式如图 5.4 所示。

```
0 qid:0 1:0 2:0 3:0.5 4:0.004413
0 qid:0 1:0.211298 2:0 3:0.5 4:0.004413
1 qid:0 1:0.53338 2:0.5 3:1 4:1
0 qid:0 1:0 2:0 3:0.25 4:0.000007
0 qid:0 1:0.137301 2:0 3:0.5 4:0.004189
0 qid:0 1:0.138519 2:0 3:0.5 4:0.000007
0 qid:0 1:0.053245 2:0 3:0.5 4:0
0 qid:0 1:0.218601 2:0 3:0.333334 4:0.009696
0 qid:0 1:0.247368 2:0 3:0.25 4:0
0 qid:0 1:0.230057 2:0 3:0.5 4:0.056504
0 qid:0 1:0.204191 2:0 3:0.5 4:0.000007
0 qid:0 1:0 2:0 3:0.5 4:0.004413
0 qid:0 1:0.211298 2:0 3:0.5 4:0.004413
0 qid:0 1:0.53338 2:0.5 3:1 4:0.990924
0 qid:0 1:0 2:0 3:0.25 4:0.000007
0 qid:0 1:0.218601 2:0 3:0.333334 4:0.010871
0 qid:0 1:0 2:0 3:0.5 4:0.004413
```

图 5.4 训练 Ranking SVM 模型的数据格式

5.4 实验与结果分析

5.4.1 实验数据

本章采用 NLPCC2016 开放域知识库问答系统评测数据集进行实验，该数据集提供了 14609 个问答对作为训练集，9870 个问句作为测试集。另外，由于该评测是基于知识库的，因此评测方还提供了一个结构化知识库和知识库中的实体集合，该知识库的大小为 2.3G，实体集合的大小为 391M。训练问句-答案对如图 5.5 所示，结构化知识库如图 5.6 所示：

```
<question id=1> 《机械设计基础》这本书的作者是谁？
<answer id=1>   杨可桢  程光蕴  李仲生
==================================================
<question id=2> 《高等数学》是哪个出版社出版的？
<answer id=2>   武汉大学出版社
==================================================
<question id=3> 《线性代数》这本书的出版时间是什么？
<answer id=3>   2013-12-30
==================================================
<question id=4> 安德烈是哪个国家的人呢？
<answer id=4>   摩纳哥
==================================================
<question id=5> 《线性代数》的isbn码是什么？
<answer id=5>   978-7-111-36843-4
==================================================
<question id=6> 《高等数学一（微积分）》是哪一门课的通用教材？
<answer id=6>   高等数学一（微积分）
==================================================
```

图 5.5 训练问句-答案对

```
万家灯火(林兆华李六乙导演话剧) ||| 别名 ||| 万家灯火↓
万家灯火(林兆华李六乙导演话剧) ||| 中文名 ||| 万家灯火↓
万家灯火(林兆华李六乙导演话剧) ||| 导演 ||| 林兆华、李六乙↓
万家灯火(林兆华李六乙导演话剧) ||| 编剧 ||| 李云龙↓
万家灯火(林兆华李六乙导演话剧) ||| 制作 ||| 北京人民艺术剧院↓
万家灯火(林兆华李六乙导演话剧) ||| 上映时间 ||| 2002年↓
万家灯火(林兆华李六乙导演话剧) ||| 主演 ||| 宋丹丹、濮存昕、米铁增、何冰↓
紫屋魔恋 ||| 别名 ||| 紫屋魔恋↓
紫屋魔恋 ||| 中文名 ||| 紫屋魔恋↓
紫屋魔恋 ||| 制片人 ||| 彼得.古伯↓
紫屋魔恋 ||| 主演 ||| 杰克.尼科尔森↓
紫屋魔恋 ||| 片长 ||| 121分钟↓
美丽的日子(王心凌演唱专辑) ||| 别名 ||| 美丽的日子↓
美丽的日子(王心凌演唱专辑) ||| 中文名 ||| 美丽的日子↓
美丽的日子(王心凌演唱专辑) ||| 发行时间 ||| 2009年11月13日↓
美丽的日子(王心凌演唱专辑) ||| 地区 ||| 台湾↓
美丽的日子(王心凌演唱专辑) ||| 语言 ||| 普通话↓
美丽的日子(王心凌演唱专辑) ||| 歌手 ||| 王心凌↓
美丽的日子(王心凌演唱专辑) ||| 音乐风格 ||| 流行↓
```

图 5.6 结构化知识库

本章分别抽取了一万条训练问句和 9870 条测试问句的候选三元组，则它们对应的候选三元组数目如表 5.2 所示。

表 5.2 候选三元组数

问句集	候选三元组数
训练问句	630960
测试问句	356773

5.4.2 数据预处理

评测方提供的问答系统的数据集中存在一些噪音词，这些词对研究问答系

统没有实际的意义，且删除它们并不影响问句本身的意思，例如“来着”、“我很好奇”等。因此，在处理数据之前，首先删除这些词。本章采用规则匹配的方式提取问句的核心部分，且这些规则是按顺序匹配的。规则表达式如表 5.3 所示：

表 5.3 规则表达式

规则表达式
(啊|呀|(你知道)?吗|呢)?(?|\?)*$
来着$
^呃(……)?
^请问(一下|你知道)?
^(那么|什么是|我想知道|我很好奇|有谁了解|问一下|请问你知道|谁能告诉我一下)
^((谁|(请|麻烦)?你|请)?(能|可以)?告诉我)
^((我想(问|请教)一下)，?)
^((有人|谁|你|你们|有谁|大家)(记得|知道)

5.4.3 实验工具

（1） 本章采用中科院分词系统对问句和知识库中的候选三元组进行分词。

（2） 在使用依存关系提取问句谓词时，本文采用 lib 调用 LTP 平台对已分过词的问句进行依存分析。

（3） 本章采用 weka 工具[13]中集成的随机森林算法来检测问句中的话题短语。

（4） 在进行余弦相似度计算时，词向量是由 Google 在 2013 年开源的 Worde2Vec 工具训练而来的，其中训练语料来自 60 年的人民日报，向量维度为 200 维，采用 CBOW 模型。

（5） 本章采用 H.Joachims 在 2006 年开源的工具 svm_rank[14]对后候选答案进行排序。

5.4.4 问答系统评测指标

由于经过候选答案排序后，返回给每个测试问句的答案是一个列表。因此本章采用平均确率（AvgP）、平均召回率（AvgR）和平均 F1 值（AvgF1）来评价该问答系统。设问句集为,其中表示第 i 个问句，N 表示测试问句数。系统返回给第 i 个问句的答案列表为，评测官方提供的第 i 个问句的标准答案列表为。

[13]https://www.cs.waikato.ac.nz/ml/weka/downloading.html
[14]http://download.joachims.org/svm_rank/current/svm_rank_windows.zip

则其评测标准的定义如公式（5.9）～（5.12）所示：

(5.9)

AvgR (5.10)

AvgF1 (5.11)

(5.12)

其中表示系统返回给的答案列表中包含的答案数，表示的标准答案列表中包含的答案数，是一个指示函数，若和中至少有一个相同的答案，则取值就是1，否则，取值为0。表示既在答案列表中，又在标准答案中的答案数目。表示第i个问句的F1值。

5.4.5 实验结果与分析

为了验证问句分类的性能对问答系统的作用，本章将实验结果与李浩[61]的结果进行了对比。本章将第四章中基于Bi-LSTM的问句分类方法应用到问答系统的答案排序阶段，来提取分类特征，并结合问句谓词和候选三元组间的余弦相似度特征、编辑距离特征和共现特征，在知识库中查找问句的答案。李浩[61]利用最大熵模型来提取分类特征且将其应用到答案排序阶段，并结合相似度特征、编辑距离特征和共现特征，在知识库中查找问句的答案。李浩[61]在计算相似度特征时，采用了三种方法：义原向量、Word2Vec、义原向量和Word2Vec。而为了观察问句分类的性能对问答系统的影响，必须保证两个实验的其他三个特征是一致的，因此，我们选择与李浩[61]的基于Word2Vec方法的相似度计算应用到问答系统的结果作对比。如表5.4所示：

表5.4 不同特征组合对问答系统的影响

特征	AvgF1(%)	AvgP(%)	AvgR(%)
相似度+共现	73.21	71.67	77.77
相似度+共现+编辑距离	75.13	73.49	80.20
相似度+共现+编辑距离+最大熵分类[61]	74.12	72.35	79.52
相似度+共现+编辑距离+Bi-LSTM分类	**76.13**	**74.49**	**83.20**

由表5.4可以看出，采用本文的分类方法提取分类特征应用到问答系统中，

平均 F1 值、召回率、准确率都有一定的提升，且平均 F1 值提升了 2.01 个百分点。这是因为基于 Bi-LSTM 模型的问句分类的准确率高于最大熵模型，使得问答系统的性能有所提高。这说明了提升问句分类的性能有助于提高问答系统的性能，且在其他条件一致的情况下，好的问句分类结果对问答系统的影响较为明显。

由表 5.4 可知，添加编辑距离特征和分类特征都能使问答系统的性能有一定的提升，分别提升了 1.92 个百分点和 1 个百分点。这说明问句谓词和候选三元组的谓语间的差异性对查找正确答案具有重要的作用，经过插入、删除、修改这三种编辑操作，将具有相同概念不同表达形式的问句谓词转化为候选三元组的谓语，缩短了问句谓词和正确答案三元组的谓语间的差异，有助于快速定位到正确答案，提升查找正确答案的准确率。所以编辑距离特征对提升问答系统的性能具有一定的作用。而加上分类特征又使问答系统的 F1 值提高了 1 个百分点。这说明研究分类特征对问答系统也具有重要的作用。这是因为给定问句的类型，正确答案的类型也就随之确定，在候选答案三元组中与问句具有相同类型的三元组的宾语越有可能成为正确答案，这样可以在候选答案集中快速定位到包含正确答案的三元组，缩短查找正确答案的时间，从而提升问答系统的性能。

5.5 总结

本章主要介绍了在知识库中查找答案的实现过程，共分成了三步：问句话题短语检测、候选答案抽取和答案排序。接下来，分别针对这三步的实现过程具体介绍。在问句话题短语检测阶段，为了检测话题短语，本阶段构建了一个训练集，并且利用构建好的训练集提取特征来训练随机森林模型，最后根据训练好的模型进行话题短语检测。在候选答案抽取阶段，将检测出的话题短语在知识库中匹配候选三元组的主语，抽取出所有的主语是话题短语的三元组。在答案排序阶段，将第四章的分类模型应用到该阶段提取分类特征，结合其他三个特征：相似度、共现、编辑距离训练排序模型，利用训练好的模型对候选三元组排序，返回排列在前面的答案列表。最后，根据问答系统的评测指标对问答系统的结果进行评价，实验结果表明，相比最大熵模型的问句分类方法，本文提出的 Bi-LSTM 的分类方法使问答系统的平均 F1 值提升了 2.01 个百分点，这说明好的问句分类性能对提升问答系统的性能具有重要作用。

6 总结和展望

6.1 总结

中文问句分析作为问答系统的第一步，为后续问答系统的实现奠定了良好的基础。而中文问句分析中关键的部分就是问句分类。它可以缩减查找正确答案的时间，对提升问答系统的性能起着一定的作用。因此，本文着重对中文问句分类的方法进行了研究。我们采用两种方法对中文问句分类进行研究：基于最大熵模型的问句分类和基于 Bi-LSTM 的问句分类研究。

基于最大熵模型的问句分类方法分别提取问句的词袋、词和词性、词向量和依存关系特征对问句进行分类，虽然实验结果表明这些特征虽然都对提升问句分类的性能起到了一定的作用，其中词向量特征与其他三个特征比较起来相对较好，但是它们都需要人工制定规则，具有一定的主观性，且这些特征的提取通常会借助分词、依存分析等自然语言处理的任务，而这些任务的准确率又会影响到最终问句分类的性能。因此，本文提出了基于 Bi-LSTM 的问句分类方法，该方法融合词、词性和位置信息生成词语的嵌入表示，然后利用 Bi-LSTM 模型自主学习问句的语义信息。该方法克服了传统方法的不足，在不需要任何繁琐规则的情况下，使问句分类的准确率达到 92.38%，比传统的最大熵模型提升了 3.63 个百分点。

为了验证中文问句分类在问答系统中的作用，本文将基于 Bi-LSTM 的问句分类方法应用到知识库问答系统的答案排序阶段，然后结合相似度特征、共现特征、编辑距离特征，利用 Ranking SVM 算法对候选答案三元组进行排序。实验结果表明问答系统的平均 F1 值达到了 76.13%，平均准确率达到 74.49%，平均召回率达到 83.20%，与采用传统的最大熵模型的问句分类方法来提取答案排序阶段的分类特征相比，平均 F1 值提高了 2.01 个百分点。

6.2 展望

虽然基于 Bi-LSTM 的问句分类方法提升了中文问句分类的准确率，但是从实验结果来看，描述类和实体类的分类精度还需要进一步提高。这是因为训练 Bi-LSTM 模型需要大量的语料，更多的语料可以更好地训练模型，而本文中描

述类和实体类的语料相对较少。因此，针对该问题，除了收集和标记更多的语料之外，还可以将本文的特征应用到多种深度学习方法中，结合多种深度学习方法的优点，更全面地理解语义，这将是中文问句分类工作下一步研究的重点。

对于知识库问答系统，答案排序阶段是提升问答系统性能最关键的一步。本文需要借助其他自然语言处理任务人工提取答案排序阶段的特征，而这些自然语言处理任务的准确率势必会影响到问答系统的性能。因此，针对此问题，利用深度学习模型自主学习答案排序阶段的特征，并对答案进行排序成为知识库问答系统下一步研究的重点。

参考文献

[1] 郑实福, 刘挺, 秦兵, 等. 自动问答综述[J]. 中文信息学报, 2002, 16(6): 47-53.

[2] Lehnert W G. A conceptual theory of question answering[C]//Proceedings of the 5th international joint conference on Artificial intelligence-Volume 1. Morgan Kaufmann Publishers Inc., 1977: 158-164.

[3] Kuo J J, Lin K K, Chen H H, et al. Question type classification and its application to a question answering system[C]//Systems, Man and Cybernetics, 2002 IEEE International Conference on. IEEE, 2002, 1: 641-646.

[4] Li X, Roth D. Learning question classifiers[C]//Proceedings of the 19th international conference on Computational linguistics-Volume 1. Association for Computational Linguistics, 2002: 1-7.

[5] Zhang D, Lee W S. Question classification using support vector machines[C]//Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval. ACM, 2003: 26-32.

[6] Hovy E, Gerber L, Hermjakob U, et al. Toward semantics-based answer pinpointing[C]//Proceedings of the first international conference on Human language technology research. Association for Computational Linguistics, 2001: 1-7.

[7] Voorhees E M. The TREC-8 Question Answering Track Report[C]//Trec. 1999, 99: 77-82.

[8] Voorhees E M, Tice D M. Overview of the TREC-9 Question Answering Track[C]//TREC. 2000.

[9] Ellen M. Overview of the TREC 2002 question answering track[C]//Proceeding of the The Text REtrieval Conference (TREC), 2002. 2002.

[10] Shen D, Lapata M. Using semantic roles to improve question answering[C]//Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL). 2007.

[11] Huang Z, Thint M, Qin Z. Question classification using head words and their hypernyms[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2008: 927-936.

[12] Al Chalabi H M, Ray S K, Shaalan K. Question classification for Arabic question answering systems[C]//Information and Communication Technology Research (ICTRC), 2015 International Conference on. IEEE, 2015: 310-313.

[13] 余正涛, 樊孝忠, 郭剑毅.基于支持向量机的汉语问句分类[J]. 华南理工大学学报 (自然科学版),2005, 33(9):25-29.

[14] 董才正, 刘柏嵩.面向问答社区的中文问题分类[J].计算机应用,2016,36(4):1060-1065.

[15] 文勖, 张宇, 刘挺, 马金山.基于句法结构分析的中文问题分类[J]. 中文信息学报,2006, 20(2): 35-41

[16] Magnini B, Negri M, Prevete R, et al. Mining Knowledge from Repeated Co-Occurrences:

DIOGENE at TREC 2002[C]//TREC. 2002.

[17]Yang H, Chua T S, Wang S, et al. Structured use of external knowledge for event-based open domain question answering[C]//Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval. ACM, 2003: 33-40.

[18]樊孝忠, 李宏乔, 李良富等.银行领域汉语自动问答系统 BAQS 的研究与实现[J].北京理工大学学报,2004(6):528-532.

[19]辛霄.面向真实环境的金融问答系统[M].哈尔滨:哈尔滨工业大学,2009

[20]贾君枝, 王永芳, 李婷.面向农民的问答系统问句处理研究[J].现代图书情报技术,2010,5:35-40

[21]任梦菲, 王鹏, 蔡恒进, 等. 股票领域中的一种中文问句分类方法[J]. Computer Science and Application, 2011, 1: 134.

[22]Lewis, David D. An evaluation of phrasal and clustered representations on a text categorization task[C]// Proc. International ACM SIGIR Conference on Research and Development in Information Retrieval. 1992:37-50.

[23]Post M, Bergsma S. Explicit and implicit syntactic features for text classification[C]//Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). 2013, 2: 866-872.

[24]Cover T M, Thomas J A. Elements of information theory[M]. John Wiley & Sons, 2012.

[25]Ng A Y. Feature selection, L1 vs. L2 regularization, and rotational invariance[C]//Proceedings of the twenty-first international conference on Machine learning. ACM, 2004: 78

[26]张宇, 刘挺, 文勖. 基于改进贝叶斯模型的问题分类[J]. 中文信息学报, 2005, 19(2): 101-106.

[27]田卫东, 高艳影, 祖永亮. 基于自学习规则和改进贝叶斯结合的问题分类[J]. 计算机应用研究, 2010, 27(8):2869-2871.

[28]孙景广, 蔡东风, 吕德新等. 基于知网的中文问题自动分类[J]. 中文信息学报, 2007, 21(1):90-95

[29]李鑫, 杜永萍, 黄萱菁, 等. 基于句法信息和语义信息的问题分类[C]//NCIRCS2004 第一届全国信息检索与内容安全学术会议论文集. 2004.

[30]Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[C]//Advances in neural information processing systems. 2012: 1097-1105.

[31]Graves A, Mohamed A, Hinton G. Speech recognition with deep recurrent neural networks[C]//Acoustics, speech and signal processing (icassp), 2013 ieee international conference on. IEEE, 2013: 6645-6649.

[32]Bengio Y, Ducharme R, Vincent P, et al. A neural probabilistic language model[J]. Journal of machine learning research, 2003, 3(Feb): 1137-1155.

[33]Xiao G, Mo J, Chow E, et al. Multi-Task CNN for Classification of Chinese Legal Questions[C]//e-Business Engineering (ICEBE), 2017 IEEE 14th International Conference

on. IEEE, 2017: 84-90.

[34]Kim Y. Convolutional neural networks for sentence classification[J]. arXiv preprint arXiv:1408.5882, 2014..

[35]张栋, 李寿山, 王晶晶. 基于问题与答案联合表示学习的半监督问题分类方法[J]. 中文信息学报, 2017, 31(1): 1-7.

[36]李超, 柴玉梅, 南晓斐, 等. 基于深度学习的问题分类方法研究[J]. 计算机科学, 2016, 43(12): 115-119.

[37]周鑫鹏.基于深度学习的问题分类的研究[D].哈尔滨工业大学,2016.

[38]徐健, 张栋, 李寿山, 等. 基于双语信息的问题分类方法研究[J]. 中文信息学报, 2017, 31(5): 171-177.

[39]Xiao G, Chow E, Chen H, et al. Chinese Questions Classification in the Law Domain[C]//e-Business Engineering (ICEBE), 2017 IEEE 14th International Conference on. IEEE, 2017: 214-219.

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

[41]Pang B, Lee L. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts[C]//Proceedings of the 42nd annual meeting on Association for Computational Linguistics. Association for Computational Linguistics, 2004: 271.

[42]Auer S, Bizer C, Kobilarov G, et al. Dbpedia: A nucleus for a web of open data[M]//The semantic web. Springer, Berlin, Heidelberg, 2007: 722-735.

[43]Bollacker K, Evans C, Paritosh P, et al. Freebase: a collaboratively created graph database for structuring human knowledge[C]//Proceedings of the 2008 ACM SIGMOD international conference on Management of data. AcM, 2008: 1247-1250.

[44]Hoffart J, Suchanek F M, Berberich K, et al. YAGO2: exploring and querying world knowledge in time, space, context, and many languages[C]//Proceedings of the 20th international conference companion on World wide web. ACM, 2011: 229-232.

[45]Cai Q, Yates A. Large-scale semantic parsing via schema matching and lexicon extension[C]//Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2013, 1: 423-433.

[46]Liang P, Jordan M I, Klein D. Learning dependency-based compositional semantics[J]. Computational Linguistics, 2013, 39(2): 389-446.

[47]Kwiatkowski T, Choi E, Artzi Y, et al. Scaling semantic parsers with on-the-fly ontology matching[C]//Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. 2013: 1545-1556.

[48]Berant J, Liang P. Semantic parsing via paraphrasing[C]//Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2014, 1: 1415-1425.

[49]Lai Y, Lin Y, Chen J, et al. Open domain question answering system based on knowledge base[M]//Natural Language Understanding and Intelligent Applications. Springer, Cham, 2016: 722-733.

[50]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 (Volume 1: Long Papers). 2014, 1: 956-966.

[51]Bast H, Haussmann E. More accurate question answering on freebase[C]//Proceedings of the 24th ACM International on Conference on Information and Knowledge Management. ACM, 2015: 1431-1440.

[52]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 (Volume 1: Long Papers). 2015, 1: 260-269.

[53]Bordes A, Weston J, Usunier N. Open Question Answering with Weakly Supervised Embedding Models[J]. 2014, 8724:165-180.

[54]Yih W T, He X, Meek C. Semantic Parsing for Single-Relation Question Answering[C]// Meeting of the Association for Computational Linguistics. 2014:643-648.

[55]Nigam K. Using maximum entropy for text classification[C]// IJCAI-99 Workshop on Machine Learning for Information filtering. 1999:61--67.

[56]Darroch J N. Generalized iterative scaling for log-linear models[J]. Annals of Mathematical Statistics, 1972, 43(5):1470-1480.

[57]宗成庆. 统计自然语言处理[M]. 清华大学出版社, 2013

[58]Vapnic. 统计学习理论[M]. 电子工业出版社, 2015

[59]严蔚敏, 吴伟民等. 数据结构（c 语言版）[M]. 清华大学出版社，2011

[60]Joachims T. Training linear SVMs in linear time[C]// ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2006:217-226.

[61]李浩. 词语相似度计算及其在问答系统中的应用研究[D], 郑州大学, 2017

个人履历、在校期间发表的学术论文以及参与项目

个人履历

张倩，女，1991 年 9 月 10 日生，河南省沈丘县人
2015 年 7 月毕业于郑州大学，信息工程学院，计算机科学与技术专业，获得工学学士学位；
2015 年 9 月至今就读于郑州大学，信息工程学院，软件工程专业，攻读工学硕士学位。

在校期间发表的学术论文

[1] 张倩,穆玲玲,张坤丽,昝红英.基于 Bi-LSTM 的问句分类方法.第十九届汉语词汇语义学国际研讨会（CLSW2018）,台湾嘉义,2018.5.26-2018.5.28

参与项目

[1] 汉语料库词义标注测试与加工过程，973 课题子任务(2014CB340504)
[2] 现代汉语语义词典语义类修订，重点实验室开放项目(北京大学 计算语言学教育部重点实验室 KLCL201401)
[3] 软件著作权：词义标注系统 v1.0(2017SR140620)

致谢

光阴似箭，不知不觉三年的研究生生涯即将接近尾声，三年的研究生生活带给了我很多的感动。走在校园中，回首过往，以前发生的点点滴滴仿佛就在眼前，心中不免多了些许的不舍。三年来，感谢陪我度过美好时光的老师、朋友和各位兄弟姐妹，正是你们的帮助和指导，才让我克服重重困难，顺利完成学业。

首先，感谢我的导师穆玲玲副教授对我生活和学业上无微不至的关怀和帮助。从课题的选择到论文的最终完成，穆老师始终如一地给予我耐心的指导和支持，我取得的点滴成绩都倾注了穆老师大量的心血。穆老师开阔的视野、严谨的治学态度、积极乐观的生活态度和精益求精的工作作风，深深地感染和激励着我，我为能遇到这样一位品德高尚、宽厚和善的好导师感到莫大的荣幸。老师对我的谆谆教诲和关怀，我时刻铭记于心。另外，在此还要感谢实验室的咎老师、张老师、贾老师、赵老师、柴老师和韩老师对我平时论文修改过程中提出的建议，让我收获良多，值此毕业之际，由衷地感谢各位老师对我的支持和帮助。

感谢实验室的兄弟姐妹和师弟师妹们在实验过程和论文写作过程中提供的热心帮助，感谢你们陪我度过这快乐的三年时光，你们的陪伴让我感觉充实和快乐，无论在哪，我都会记得这份感动。

最后，我要感谢我的父母对我的关心、鼓励和支持，是你们让我懂得了责任和感恩，是你们在背后默默地给我支撑和帮助，你们对我如此无私的付出，我会用一生去回报。因为你们的关心和帮助，让我觉得很幸福，在以后的工作中，我会努力做到最好，作为对你们最好的回报。