心求至善

基于本体的地理知识问答

张赏

东南大学

学校代码: 10286
分 类 号: TP311
密 级: 公开
U D C: 004.4
学 号: 153305

東南大學
硕士学位论文

基于本体的地理知识问答

研究生姓名: 张赏
导 师 姓 名: 高志强 教授

申请学位类别 工程硕士 学位授予单位 东 南 大 学
一级学科名称 计算机科学与技术 论文答辩日期 2018 年 6 月
二级学科名称 计算机技术 学位授予日期
答辩委员会主席 评 阅 人

2018 年 5 月 23 日

学校代码: 10286
分 类 号: TP311
密 级: 公开
U D C: 004.4
学 号: 153305

东南大学

SOUTHEAST UNIVERSITY

硕士学位论文

基于本体的地理知识问答

研究生姓名: 张赏

导 师 姓 名: 高志强 教授

申请学位类别 工程硕士 学位授予单位 东南大学

一级学科名称 计算机科学与技术 论文答辩日期 2018 年 6 月

二级学科名称 计算机技术 学位授予日期

答辩委员会主席 评 阅 人

2018 年 5 月 23 日

东南大学

硕士学位论文

基于本体的地理知识问答

专 业 名 称: 计算机科学与技术

研究生姓名: 张 赏

导 师 姓 名: 高 志 强 教 授

本工作受到国家 863 计划项目 2015AA015406，国家自然科学基金 61170165、61602260 和 61502095 的资助。

ONTOLOGY-BASED GEOGRAPHIC KNOWLEDGE QUESTION ANSWERING

A Thesis submitted to

Southeast University

For the Academic Degree of Master of Engineering

BY

Zhang Shang

Supervised by:

Prof. Gao Zhiqiang

Suzhou Joint Graduate School

Southeast University

2018/5/23

东南大学学位论文独创性声明

本人声明所呈交的学位论文是我个人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得东南大学或其它教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

研究生签名：_____ 日 期：_____

东南大学学位论文使用授权声明

东南大学、中国科学技术信息研究所、国家图书馆有权保留本人所送交学位论文的复印件和电子文档，可以采用影印、缩印或其他复制手段保存论文。本人电子文档的内容和纸质论文的内容相一致。除在保密期内的保密论文外，允许论文被查阅和借阅，可以公布（包括刊登）论文的全部或部分内容。论文的公布（包括刊登）授权东南大学研究生院办理。

研究生签名：_____ 导师签名：_____ 日 期：_____

摘　要

在人工智能领域，自动解答高考题是一项很具挑战的任务。与一般事实性问答的问题不同，高考题带有很强的选拔性，其问题考察形式多变，其答案求解往往不能一步得到，通常需要做进一步的知识推理。在辅助解答高考地理题时，目前面临两个问题：第一是缺乏高度结构化的地理核心知识库，第二是地理问题表达形式多样，导致问题理解困难。针对以上两个问题，本文做了如下三个工作：

（1）为解决高度结构化的地理核心知识库缺乏问题，本文构建了中文地理本体（Chinese Geographic Ontology，CGeoOnt）。该本体以人教版高中地理教科书为知识源，使用万维网本体语言 (Web Ontology Language, OWL) 为知识表示语言，以课本章节为知识体系，人工总结其核心地理概念、地理关系、地理考点，并将其表示为本体形式。同时，本文将构建的本体 CGeoOnt 与本体 Clinga 进行融合，得到一个更大规模的中文地理本体知识库。

（2）为解决地理问题问法多样导致其难以理解问题，本文使用基于注意力机制的知识库问答模型。该模型以双向长短期记忆网络为基础问答模型，结合注意力机制对地理问题、答案进行表示。答案中每个词的向量生成，均结合其对问题各词的注意力权重分配，使答案可以更好的对齐问题中相应的关键信息，减弱无效信息的干扰，因此更易区分正确答案和错误答案。实验表明，该问答模型对于辅助解答地理高考题具有很好的参考和应用价值。

（3）为解决中文地理问答模型在训练和测试中数据集缺失问题，本文从互联网收集了一个问法多样的中文地理问题集。本文使用百度问题推荐以及百度搜索 API，以本体知识库高频核心知识三元组为数据源，依次访问到二十万个 Web 地理问题，然后半自动加人工挑选出其中的有效问题，再人工从知识库寻找问题答案，形成最终地理问答数据集。

关键词： 地理高考，本体构建，知识库问答，双向长短期记忆网络，注意力机制

Abstract

In the field of artificial intelligence, it is a challenging task to automatically answer the college entrance examination (namely GaoKao) questions. Different from the questions of the general factoid question answering, the questions of GaoKao have a strong selection, the problem's forms are changeable, and the solution often cannot be obtained in one step, and it usually needs further knowledge inference. At present, there are two problems in assisting in solving the geography problem of GaoKao: the first problem is the lack of a well structured geographical core knowledge base (KB), and the second problem is that geographical problems are expressed in various forms, which makes it hard to understand. In view of the above two problems, this thesis has done the following three researches:

(1) In order to solve the problem of the lack of a well structured geo-core knowledge base, this thesis constructs a Chinese Geographical Ontology (CGeoOnt). The ontology takes the version of the geography textbook published by people's education press as knowledge source, uses the World Wide Web Ontology language (OWL) as the knowledge representation language, takes the textbook chapter as the knowledge system structure, summarizes its core geographical concept, the geographical relation, the geographical test center, and expresses them in ontology form. At the same time, this thesis integrates the constructed CGeoOnt with ontology Clinga(chinese linked geographical dataset), and obtains a more large-scale chinese geographical ontology knowledge base.

(2) In order to solve the problem of understanding geo-questions with various forms, this thesis employs the knowledge base question and answering model based on attention mechanism. The model is based on bidirectional long and short term memory (Bi-LSTM) network, combined with the attention mechanism to express the geographical problem and the answer, the vector generation of each word in the answer is combined with its attention weight distribution, so that the answer can better align the key information in the problem, weaken the interference of invalid information, which makes it easier to distinguish between correct answers and wrong answers. The experimental results show that this model has good reference and application value to assist in solving geography GaoKao questions.

(3) In order to solve the problem of the lack of dataset in the training and testing procedures of Chinese geography question and answer model, this thesis collects a variety of chinese geography problem sets from the Internet. This thesis uses the Baidu question recommendation API as well as the baidu search API, takes the the high frequency core knowledge triples in ontology knowledge base as data source, then has access to 200,000 web geography questions,

picks out the effective questions semi-automaticly and manually, then manually seeks the answers of the questions according to the knowledge base to obtain the final geography question and answering dataset.

Keywords: Geography College Entrance Examination, Ontology Construction, Knowledge Base Question Answering, Bi-LSTM, Attention Mechanism

目录

术语与符号约定

BOW	Bag Of Words
CGeoOnt	Chinese Geographical Ontology
Clinga	Chinese Linked Geographical Dataset
CNNs	Convolutional Neural Networks
KB	Knowledge Base
KBQA	Knowledge Base Question Answering
LSTM	Long And Short Term Memory
NER	Named Entity Recognition
POS	Part Of Speech
RNN	Recurrent Neural Network

第一章　绪论

1.1　研究背景

近年来，一个比较热的人工智能挑战是让计算机参加高考。早在 2011 年，日本国立情报学研究所（NII）发起了一项名为“东大机器人项目”（Todai Robot Project）的人工智能项目，其最终目的是让此“高考机器人”能够在 2021 年通过东京大学的入学考试[1]。2015 年，国家也启动了 863“基于大数据的类人智能关键技术与系统”项目，其目的为攻克高考九门学科中的四门，即语文、数学、地理、历史[2]。本文工作也是对辅助解答高考地理多选题的一些思考尝试，图 1 所示为 2016 年上海高考地理多选题:

（**2016 年上海-高考**）今年 4 月，太平洋周边某些国家出现异常高温干旱天气，有专家认为这与厄尔尼诺有关。根据厄尔尼诺影响的一般规律判断，发生干旱的国家可能是（）

A. 日本　　**B.泰国**　　C.智利　　D.秘鲁

图 1.1: 地理高考多选题举例

题中划线部分为题目问题，划线部分之前为问题的背景知识介绍。由题可知，问题考察“厄尔尼诺现象会使哪些国家或地区产生干旱现象?”，要解答此问题，计算机必须具备“厄尔尼诺现象”相关的核心知识。如此处需要知道地理知识，厄尔尼诺现象使印度、东南亚、印度尼西亚和澳大利亚产生干旱，然后根据选项中“泰国”属于东南亚，可知此题答案选“泰国”。由上述解题过程可知，解答此类地理问题需要高度结构化的地理知识，并且地理知识表示需包含丰富的语义信息。如此题需要知道“（厄尔尼诺现象，导致干旱的国家，印度、东南亚、印度尼西亚、澳大利亚）”三元组，同时也需要知道“（东南亚，包括，越南、老挝、柬埔寨、缅甸、泰国、马来西亚、新加坡、印度尼西亚、菲律宾、文莱、东帝汶）”三元组，并且还需要知道“东南亚”的类别是一些国家的集合，“泰国”的类别是东南亚。

鉴于以上分析，解答高考地理题多选题一般可粗粒度分为两步，第一步为匹配求解问题所需的知识库三元组知识，第二步为根据结果三元组知识作进一步推理得出最终答案。作为辅助解答高考地理选择题中的尝试，本文的工作集中在第一步上，即先构建解答地理高考题所需要的地理核心知识库，再从该知识库中找出最可能回答所求地理问题的三元组知识。

地理解题核心知识库需要高度结构化的知识表示，并且知识需包含丰富的语义信息，如知识类别、关系等。显然，无结构的文本文档以及半结构化的数据（如 xml、json 格式）表示形式都无法满足要求。在结构化表示领域知识时，本体可以很好对领域知识

建模，并且表示出计算机可以处理的带有丰富语义的形式化定义[3]。前期的地理知识以地理教科书形式存在，地理教科书知识分章节层次描述，计算机是无法处理此种自然语言式的语义关系。因此，需要使用本体对其建模，通过本体中的实体、类别、属性、关系等术语，描述地理中概念（如地球、星球等）的属性信息、类别信息和各概念之间的关系。地理核心知识通过三元组（主、谓、宾）形式得以更精炼的表示，地理核心概念层次关系也明显，更适合做进一步的推理。

基于构建的地理核心知识库之上，本文需要构建一个问答系统。给定一个地理问题，系统能返回求解该问题所需的地理知识三元组。目前，基于知识库的问答任务有两个主流的研究方向：基于语义解析[4;5;6;7] 和基于信息检索[8;9;10;11]。基于语义解析的方法一般先构建一个语义解析器，然后运用该语义解析器将自然语言问句转换为特定类型的逻辑表达式，如带类型的 lambda 表达式 (typed lambda calculus)、lambda 依存组合语义[12;6;13]。基于信息检索的方法通常先从知识库检索一系列候选答案，然后对问句和候选答案进行特征抽取并打分，选出得分最高的结果作为最终答案[9;14]。基于信息检索的方法更简单，实现也更灵活，在开放域知识库 Freebase 上的问答研究也表明，该方法可以达到与基于语义解析方法相近的 F 值[10;11]。随着深度学习的兴起，神经网络被运用到知识库问答中提升已有模型，基于神经网络的问答模型只需将问题和答案分别表示成低维语义向量，然后通过计算向量相似性，即可获得最相似的候选答案作为最终答案。问句和答案的向量表示是基于神经网络问答模型的一个重要环节，有些研究比较侧重答案表示，如运用候选答案在知识库子图中的重要性[9] 或者答案的类型和上下文[10]。这些研究往往使用简单的词袋模型来表示问题，忽视了问题与答案的关联性[9]。还有研究使用 Attention 机制根据不同答案的不同注意力方面来表示问题[15]，取得了比较好的效果。

分析本文搜集到的地理问题可知，地理问题表达形式多样，无效信息较多，一条地理三元组知识往往可以成为多个地理问题的答案。如三元组——“（季风气候，生产优势，夏季高温多雨、雨热同期）”，可以作为“亚热带季风气候在发展农业生产方面有什么优势”和“温带季风气候在农业生产方面的显著优势是 _ 百度知道”这两个问题的答案。虽然上述两个问题表述不一样，但其问题核心均考察“季风气候对生产的优势”，因此相对答案三元组而言它们是等效的。这也说明，在做问答时，单独的将问题和答案表示成向量是不准确的，至少是不能表示出问题和答案之间的关系。因此可以结合 Attention 机制，在答案向量表示时同时结合对问题的 Attention 权重，这样可以更合理的表示问题和答案的关系，同时答案中重要信息与问题中重要信息对齐，这样也减弱了问题中无效信息的干扰，使更容易筛选出正确答案。

1.2　研究内容

本文为辅助解答高考地理多选题所做的尝试性工作，本文核心内容为构建地理核心知识库，并从知识库中找出可以回答所求地理问题的三元组知识。因此，本文研究如何使用本体更准确、更精炼地表示地理教科书中的知识，从而构建一个高质量、高可用性

的地理本体知识库。同时，本文研究如何更准确表示形式多样的地理问题，从构建的地理知识库中找出可以回答该问题的地理知识三元组，便于解题组做进一步的答案推理，得出问题最终答案。

本文主要研究内容如下：

（1）为解决高度结构化的地理核心知识库缺乏问题，构建中文地理本体 CGeoOnt。CGeoOnt 的构建使用 OWL 本体语言，将地理教材中核心考点的概念属性、核心概念之间的关系形式化表示。然后，运用启发式规则将 CGeoOnt 与本体 Clinga 进行融合，采取人工做最终的校验，最后得到规模更大的中文地理本体知识库。

（2）为解决地理问题问法多样导致其难以理解问题，使用基于神经注意力机制的双向长短期记忆内存网络问答模型。问答模型不是独立对问题和答案进行词向量表示，而是在充分考虑问题和答案之间的依赖关系基础上，结合问题对答案进行综合词向量表示。使正确答案三元组与问题关键信息更好的对齐，减弱非关键信息的干扰，从而更易区分相近的答案三元组，使问答模型辨别能力更强。

（3）为解决中文地理问答模型在训练和测试中数据集缺乏问题，从互联网收集了一个问法多样的中文地理问题集。问题集中的问题知识均来自地理知识库中的核心出题考点，运用百度问题推荐框 API（Application Programming Interface）和问题搜索 API，从互联网获取这些考点的相关地理问题，经机器半自动筛选加人工筛选出其中的有效问题。最后，人工从本文构建的地理知识库中找出能回答这些问题的地理三元组知识，形成最终的问题、答案对数据集。

1.3 论文组织

本文共分为四章，各章的主要内容如下：

（一）第一章介绍相关的研究背景、研究内容以及论文的组织结构。

（二）第二章介绍论文本体构建、知识库问答方法的研究现状以及论文的工作特色。本体构建研究中介绍了主流的人工本体构建方法，并对比几种方法的优缺点；知识库问答方法研究中介绍了两个主流知识库问答——基于语义解析的知识库问答、基于信息检索的知识库问答的流程和优缺点；论文的工作特色中介绍了论文与之前研究的不同之处，以及论文的亮点和创新点。

（三）第三章介绍基于本体的地理知识问答系统的具体实现过程。论文先介绍论文任务和论文系统结构图，然后介绍本文地理本体构建的具体流程，再介绍在构建的地理知识库之上如何构建问答系统用于回答地理问题，最后介绍论文的实验、实验结果和本文地理数据集的构建。

（四）第四章介绍整个论文内容的总结，总结本文的工作亮点及不足，并尝试分析未来本文工作可以提升的方向。

第二章　相关研究

本章介绍本文相关研究工作，主要包括本体和问答的研究现状。首先介绍本体相关内容，包括本体核心构成要素：个体、类、属性、关系和本体描述语言；然后介绍本体构建方法研究和知识库问答的主流方法，包括基于语义解析的方法和基于信息检索的方法；最后介绍本文的工作与已有工作不同之处。

2.1　本体构成要素

从计算机科学角度看，本体是对相关领域知识的一种高度结构化、层次化的抽象建模，这种建模表示包含一系列计算机可以处理的形式化定义[3]。运用本体可以很好地表示出领域中核心概念的语义信息和概念之间的相互关系。通过 4 个核心要素个体 (Individuals)、类 (Classes)、属性 (Properties) 及关系 (Relationships)，本体能够将领域知识以一种类似现实世界的组织方式形式化地表示出来，并且从某种程度上，既符合人的直观对领域知识层次分类的理解，又适合计算机存储、检索和推理。因此，本体是一种很好的结构化知识库建模方式。

2.1.1　个体 (individuals)

个体，又叫实例 (instances)，是本体中最基本、最底层的组成单元。本体大多数描述都是企图更准确、更详细的描述出个体的信息特征。常见的人、动物、汽车、天体、星球等中的具体对象都可以看做个体，如地球、月球、太阳都是个体，就算是抽象的数字、单词等也可以视作个体。本体的一个很重要任务就是对本体领域中的个体进行层次化的分类，使得不同个体可以很好的进行区分或者是使其可以建立某种关系。

2.1.2　类 (classes)

类，又叫类型 (type)、类别 (sort)、种类 (kind) 或者类目 (category)，常常指某个个体的上层延伸或者内涵。类是一些特征相似的个体构成的一个集合，或者是由一些子类构成的大类集合。如下为类的举例：

（1）人：人包括黄种人、白种人等类型，具体的张三、李四是人的个体。

（2）动物：动物包括无脊椎动物和脊椎动物两大类。

（3）汽车：汽车表示所有具体品牌汽车的类别。

（4）天体：天体包括地球、月球、彗星、流星等宇宙空间的物质。

（5）行星：行星包括地球、金星、木星、水星、火星、土星、天王星、海王星等。

通常情况下，一个个体可以属于不同的类别，这样使表达更灵活、表达力更强。同时，一个类可以包含其它类或者被其它类包含，这样构成了类的层次关系。一个类 A 被另一类 B 包含称为：A *is subClassOf* B，通过这个关系可以得到很重要的性质，即 B 类具有的性质 A 类也同样具有。同样，一个类可以被多个类包含，也就是说一个类可以有多个父类。正是类别的上述层次关系，使知识不仅可以表示其自身的特征，还可以表示其与其它知识的关系，而且这种关系是非常接近人类的概念思维，所以知识建模很直观、易于理解。

2.1.3 属性 (Properties)

属性用于表示个体间的关系 (ObjectProperty) 或者个体与其数据值之间的关系 (Datatype-Property)。例如人相关的属性 *hasWife*、*hasHeight* 和 *hasAge*，*hasWife* 表示两个人（两个具体的人是两个个体）之间是夫妻关系，*hasHeight* 和 *hasAge* 表示一个人的身高和年龄，身高、年龄值为数值类型。同样，属性与类结构类似，具有子属性层次结构。

属性的另外一个重要特点是：属性含有定义域 (domain)、值域 (range) 的限制 (restriction)。运用属性的这一限制，可以对属性两边的个体做相关的类别推理。如根据申明两个个体是 *hasWife* 关系，可以推导出这两个个体的类别都是人，且更进一步该关系左边的个体类别为男人，右边的个体类别为女人。

2.1.4 关系 (Relationships)

关系用来表示对象之间是以怎样的方式相联系在一起的。以汽车系列举例：“福特探险者是福特野马的下一代”，此例子体现出“福特探险者”和“福特野马”这两个对象存在着“下一代”的关系，这一事实可以表示为：

“福特探险者 *is defined as a successor of* 福特野马”

此关系表达出“福特探险者系列”取代了“福特野马系列”这一事实，显然这种关系是存在着方向的。同样可以用此关系的反向关系，即“上一代”来表示上面的事实——“福特野马是福特探险者的上一代”。关系的总集合就构成了领域本体的丰富语义信息，因此关系的表达能力大小很大程度决定着本体对领域的抽象建模能力。如下介绍两种重要的关系：

（1）包含关系（subsumption relation）

包含关系主要有 *is-a-superclass-of*、*is-a-subclass-of* 和 *is-a-subtype-of*，分别表示父类关系、子类关系和子类型关系。这些关系都是表达一种上下位的关系，特别是其中的 *is-a-subclass-of* 关系，它体现出一种很强的分类学思想，可以直观地对领域概念进行分类，并且表示出这些类别的层次关系。

（2）总分学关系（mereology relation）

总分学关系指的是一种部分 (*part-of*) 与整体的关系，表示一个对象是另一个复合对象的一部分。还是以“福特探险者”系列为例，“方向盘 *is-a-part-of* 福特探险者”，表示方向盘是福特探险者汽车的一个部件。

除了包含关系和总分学关系以外，本体中还有一些其它的关系，这些关系不一定表示层次关系，其往往是该本体领域中的特定业务关系。这种领域的特定关系被用来表达领域独特的事实知识，构成了自身领域本体的特色，因此不同领域本体表示往往差别比较明显。

2.2　本体描述语言

使用本体描述语言可以很好的对领域本体进行层次化的表示。常见的本体描述语言有：Ontolingua、OCML、OKBC、FLogic、LOOM、DAML、SHOE, OIL、XOL、XML、RDF、RDFS、OWL[16]。其中，由 W3C 推荐的 XML、RDF、RDFS 以及 OWL 使用最为广泛。

XML(Extensible Markup Language)[17] 是一种标记语言，通过其标记可以对结构化文档进行分层的语法表示，并且易于机器处理和人类阅读。然而，XML 标记缺乏对文档的含义进行约束，标记内部也缺乏结构化定义，因此很难充分描述出本体中的四个常见核心要素。RDF(Resource Description Framework)[18] 是一种描述对象（资源）以及对象之间关系的图数据模型，其兼容 XML 语法，并且含有简单的语义。RDFS(RDF Schema)[19] 是扩展的 RDF 词汇表，这里的词汇表指定义为个体、类、属性和关系的术语名称。RDFS 通过扩展了 RDF 中没有的属性和类层次结构语义，也即通过定义子属性（subPropertyOf）、子类（subClassOf）、属性定义域约束、属性值域约束来增强描述资源的表达能力。尽管 RDFS 相对 RDF 的描述资源能力更强，RDFS 仍然是一种相对简单的本体语言，其描述资源能力依然很有限[20]。例如：RDFS 无法描述类的不相交关系，如类“男人”和“女人”是不相交的，但其只能描述“男人”和“女人”同属于“人”；同时，RDFS 也无法描述类的布尔组合（并集、交集、补集）关系，如“人”无法定义成“男人”和“女人”的并集等。为弥补 RDFS 表达能力的不足，W3C 又推出了表达能力更强且具备强推理能力的本体语言——OWL(Web Ontology Language)[21]，OWL 定义了逻辑类的关系表示，即提供了针对逻辑与、或、非的领域关系表示，可以有效的表示类的并集、交集、补集运算，因而可以表达更复杂的本体知识。

OWL 的一个很重要设计思想是在知识表达能力和推理效率之间找到一个平衡。因此，在其不同的表达能力和推理效率设计中, 为满足不同用户对本体建模需求，又诞生了三个子语言，即 OWL Lite、OWL DL(Description Logic) 和 OWL Full。这三个子语言描述能力依次增强，推理复杂度逐渐提高。OWL Lite 更关注本体表达的简洁性，其表达能力相对其它两种语言较弱，但其推理最高效，因此 OWL Lite 更适合于对表达能力要求不是太强的领域；OWL DL 比 OWL Lite 表达能力更强，与 OWL Full 比具有计算完备性（所有结论均可计算）和可判定性（有限时间内所有计算均可终止），同时 OWL DL

支持有效的推理，因此在既对本体语言表达能力要求高，又保证推理可判定性时，可以选择此语言；OWL Full 的表达能力最强，正因其表达更灵活、约束较少，也使其推理不可判定，但 OWL Full 有完全兼容 RDF 的优点，这也是前两种语言不具备的，因此在以兼容 RDF 为主要建模目标的场景，OWL Full 语言是较好的选择。

2.3 本体构建方法研究

目前，本体构建方法[22] 主要有两种：第一种是在领域专家的指导下，使用本体描述语言表示领域本体；第二种是从结构化、半结构或者无结构数据抽取本体要素，形成领域本体。第一种本体构建方法往往采用纯人工方法，由于人的主观性，不同领域专家构建出来的本体常常相差很大且构建效率低。但从局部来看，专家构建的本体知识质量很高，专家站在领域高度对知识进行了专业化的总结、提炼，知识表达更准确、合理。所以，在对知识表示专业程度要求很高、知识数量较小的领域（如高考地理解题），此方法可以达到很好的效果。第二种本体构建方法是为了缓解第一种方法中的人为主观性和低效性而提出的，其使用自动或者半自动的方法构建本体，既节省时间又提高了知识表示的一致性。该方法很大程度依赖于自动、半自动技术的能力，构建出来的本体往往噪音较多、抽象程度不高（知识没有经过深度提炼、总结）。因此，在一些对知识准确度要求很高的场景，如地理高考解题本体构建等，此方法不是很适合，但是在一些对噪音容忍度比较高的场景，如通用领域的本体构建，此方法运用很广且效率很高。

由于本文构建用于辅助解答高考地理多选题的本体，其知识要求精炼、准确、少噪音且知识量不大（只包含核心地理高考考点），因此需要地理领域相关专业人士人员构建，保证质量。本文也只详述人工本体构建的研究现状，对于自动和半自动本体构建方式的研究现状，本文不加以叙述。

国内外常见人工本体构建方法有：IDEF5 法[23]、骨架法[24]、TOVE 法[25]、METHONTOLOGY 法[26]、KACTUSK 工程法[27;28]、SENSUS 法[29] 以及七步法[30] 等。IDEF5 法、骨架法及 TOVE 法常用于企业领域本体构建。IDEF5 法使用图表以及很细化的说明形式来获取企业业务存在的概念、属性及关系，从而形成本体；骨架法是一种流程图导向的本体构建方法，其描述的是一种本体构建的方法框架；TOVE 法是通过本体建立企业知识的逻辑（一阶逻辑）模型。其它四种方法用于构建领域知识本体。METHONTOLOGY 法是在构建化学元素周期表本体基础上发展而形成的通用本体构建方法，此方法偏向软件工程的思想，本体的构建也引入了管理、开发和维护三个很具软件工程特征的阶段；KACTUS 法是对应用驱动的本体开发方法，侧重对已有本体的复用和扩充方法；SENSUS 法提供一种自顶向下的层级结构本体构建方法，偏向操作性指导。七步法是基于本体工具 protege[1] 的构建方法，主要分七个步骤构建本体，比较偏向实用性、操作性。具体七种本体构建方法比较如表 2.1 所示。

[1] https://protege.stanford.edu/

表 2.1：七种本体构建方法比较

名称	构建方式	方法细节	是否支持演进	生命周期	应用领域	与 IEEE 标准一致性
IEDF5	人工	详细	否	无	企业	不完全一致
骨架法	人工	简单	否	非真正的生命周期	企业	不完全一致
TOVE 法	人工	简单	否	非真正的生命周期	企业	不完全一致
METHONTOLOGY 法	人工	详细	否	有	化学	不完全一致
KACTUSK 法	不确定	简单	否	无	电子网络开发	不完全一致
SENSUS 法	不确定	一般	否	无	电子科学	不完全一致
七步法	半自动	详细	否	非真正的生命周期	医学	不完全一致

分析以上七种人工本体构建方法可知，这些构建方法都缺乏对本体进化的考虑，本体只立足当下情况，没有考虑后期本体的更新，也使本体知识不能与时俱进。同时，这些方法人工参与力度很大，使用的技术较简单，构建时并没有统一的标准规范对其指导，构建人员均需从自身领域特点出发进行扩展与缩减。因此，本体构建方法相差还是比较大，这也表明统一的本体构建规范、评价标准还不成熟。

2.4　问答方法研究

问答是人工智能领域中的一个热门的研究问题，它综合运用了各种自然语言处理技术。对于用户提出的某一个问题，问答系统往往可以给出简短、精准的答案组织形式，而非一系列的相关网页文档供用户参考，省去了用户额外从大量的相关网页文档中寻找所需确切信息的时间[31][32]。同时，问答技术集成了定位、抽取以及表示出针对用户提出的自然语言问题答案的丰富功能，因此受到广泛的关注[33][34]。

问答一般分为两类：开放域的问答和封闭域的问答。开放域的问答又叫无结构数据的问答，一般是从开放的网页或者文档中抽取所需的问题信息。封闭域的问答又叫受限域的问答或结构化数据的问答，其往往需要预先定义的知识源，例如领域本体或关系数据库管理系统[35][36]。查询数据库的方式有两种，第一种为结构化查询，如 SQL，另一种为自然语言查询，即用户用自然语言组织其问题，而不需要一些专业术语的限制[37]。结构化查询虽其功能强大，但不易使用，对缺乏训练的普通用户不友好。相反，自然语言查询方式对用户更友好，用户可以轻松组织自然语言进行复杂的问题查询。

对于知识问答而言，其强调如何根据给定问题作出精准回复，其答案要求不能包含无效的信息。因此，知识问答的知识源常常选取高度结构化、包含丰富语义信息的数据，如本体等。建立于高度结构化数据之上的问答也叫做知识库问题，本文将详述知识库问答的主流研究方向和方法。

目前，知识库问答 (knowledge base question answering，KBQA) 主要有两个主流的研究方法：基于语义解析、基于信息检索。分别如下介绍：

2.4.1 基于语义解析的 KBQA

基于语义解析的方法一般先构建一个语义解析器，然后运用该语义解析器将自然语言问句转换为特定类型的逻辑表达式，如带类型的 lambda 表达式 (typed lambda calculus)、lambda 依存组合语义。以问题“姚明的老婆出生在哪里？”为例，问题经过构建的语义解析器解析过后，可以表示为如下 lambda 表达式：

$$\lambda x.\ 配偶\ (姚明,\ y) \wedge 出生地\ (y,\ x)$$

在此 lambda 表达式中，λ.x 表示该表达式的变量 x，关系配偶 (姚明, y) 表示姚明的配偶（也即问题中的老婆）y，关系出生地 (y, x) 表示 y 的出生地是 x，$\wedge$ 表示上述两个关系同时满足，整个句子表达的含义也就是“姚明的配偶的出生地”，即为题中问题“姚明的老婆出生在哪里？”的一种较正式表达。

传统的语义解析器需要有人工标注的逻辑表达形式作为监督知识，并且它们是在特定领域（受限域）进行相关操作，逻辑谓词的数量也较少[38][12][39]。Zettlemoyer 和 Collins 在研究将美国地理领域问题（Geo880[2]中问题）表示成 lambda 表达式时，使用人工标注的问题和对应逻辑表达式作为训练数据，去学习语义解析器，而且研究中定义的领域谓词也很有限，如类型、大小、多少、边境、城市、州、河流等。

由于获取人工标注的逻辑表达式进行有监督的训练语义解析器代价太大，并且人工标注的量始终有限，因此有相关研究尝试在不需要人工标注的逻辑表达式的前提下，学习一个语义解析器。Berant[40] 等人训练了一个适应 freebase 的语义解析器，该研究主要分两步进行：第一步是构建较完备的词汇表，第二步是桥接（Bridging）操作。构建词汇表也就是将自然语言与知识库实体或实体关系进行单点映射，也可称作对齐（alignment），自然语言实体到知识库实体的匹配较简单，可以使用一些简单的字符串匹配方式，如“Obama was also born in Honolulu.”中 Obama 映射到 Barack Obama，但是要将例句中的自然语言短语“was also born in”映射到相应的知识库实体关系 PlaceOfBirth，则较难通过字符串匹配的方式建立映射。此处作者使用一种远程监督的思想，作者使用开放信息抽取软件 ReVerbopen IE system[3]从 ClueWeb12[4]数据集上抽取 15,000,000 个三元组构成一个数据集，然后将短语对齐到知识库关系。完成词汇表的构建后，仍然有些轻动词（light verb），如 go，where，do 难以映射到知识库，比如“Which college did Obama go to?”中的“go to”无法映射到知识库关系，因此作者使用“桥接”操作，从知识库中找

[2]http://www.cs.utexas.edu/users/ml/geo.html

[3]http://reverb.cs.washington.edu/

[4]http://lemurproject.org/clueweb12/

出一个中间二元关系（本例中为 Education）来将当前实体和关系连接起来。该研究在 free917[5] 上取得了较好效果。

使用有监督的数据训练语义解析器在知识库数量较小时质量很高，但是很难适应大型知识库，如 freebase 等。Cai[5] 等人将直接根据标注的数据训练语义解析器的过程简化为三个子过程，以减小在遇到未见过测试数据时的失败率。这三个子过程分别是标准的监督学习算法、模式匹配（schema matching）和模式学习（pattern learning）。该研究应用现有的监督训练算法对有标签数据集进行语义解析，使用结构匹配技术查询自然语言词汇与本体知识库中符号的相关联性，使用模式学习技术将新的（自然语言词，本体词汇）对合并到训练的语义解析器中。此研究优点是将一个问题分解为了三个子问题，使之可以运用每个子问题领域的成熟技术。同时，三个子过程联合确定问题的逻辑表达式，比单纯的使用有监督的问题到逻辑表达式对训练，而得到的语义解析器可靠性更强。该研究在 free917 数据集比单纯的有监督学习语义解析器的算法 F 值高出 0.42%。

之前的研究在生成语义解析器时都很少运用到知识库的相关结构图信息，因此，问题中的词表达方式和知识库中术语的表达方式仍然相差很大，相互映射也比较困难。为了解决上述问题，并且充分利用知识库的结构信息，Yih[13] 等人提出了一个紧密结合知识库的语义解析器。作者定义了一个类似知识库子图的查询图（query graph），该图可以直接映射到一种 λ 演算，因此作者把问句的语义解析过程转化为查询图生成过程，并且转化为按阶段状态（stage state）和动作（action）的搜索问题，其中每个阶段状态是查询图表示中的候选解析，每个操作都定义了一种扩展关系图的方法。作者将动作分为三步，第一是识别问题中的主题实体，第二是寻找答案与主题实体之间的主要关系，第三是使用附加的、描述答案需要有的属性约束，或者是答案与问题中其它实体的关系来扩展查询关系图。这种方式通过将问句表达部分靠近知识库中的某些实体和关系，使得只需要关注最可能成为正确查询图的区域，使搜索空间大大缩小，搜索效率也大大提高。该研究在 WebQuestions[6] 上也超过了之前的方法，并取得较高的 F 值 52.5%。

2.4.2　基于信息检索的 KBQA

基于信息检索的方法一般根据问题中的关键信息去知识库查询一批候选答案，然后运用排序打分技术对候选答案进行打分，选出得分最高的候选答案。具体的操作如下：

（1）识别问题中的主题实体，即问题考察的核心实体。

（2）根据问题主题实体，从知识库中查询该实体以及其相关联的实体子图，子图的边作为候选答案集合。

（3）使用规则或者模板等，人工或者自动、半自动地构建问题的特征向量，然后根据问题特征向量对候选答案进行筛选。或者直接对问题和候选答案进行分布式的表示，即对问题、答案进行向量建模，然后根据问题、答案的向量相似性来筛选最终答案。

[5]https://github.com/pks/rebol/tree/master/data/free917

[6]https://nlp.stanford.edu/software/sempre/

Yao 和 Van[8] 首先将信息检索方法运用到知识库 freebase[7]问答上，其使用了信息抽取的技术。该研究首先使用 Standford CoreNLP[8]构造出问题的语法依存树（dependency tree），然后识别问题中的问题词（如 what、who、when 等）、问题焦点词（常暗示着答案的类型词，如 name、place 等）以及通过命名实体识别来确定主题词，通过词性标注获取问题的中心动词。再然后，根据主题词找出知识库中对应的主题词的子图，包含跟主题词相关联的实体结点边，所有实体结点的边构成问题候选答案三元组集合。最后，选取候选答案实体结点的关系、结点属性构成候选答案的特征向量，并使用问题和答案特征向量构建一个逻辑回归分类器。总体说来，此方法也运用到一些语言学知识，但总体还是较符合人的直觉。

为解决问题、答案抽取特征向量时对语言学知识和人工规则的依赖，一些研究尝试使用语义向量来对问题、答案进行分布式表示（Distributed Embedding）。Bordes[41] 等人率先使用基于神经网络的方法在开源知识库 ReVerb[42] 上进行问答任务。Bordes 等人将问题和知识库三元组都表示成低维向量，并且使用余弦相似度来计算出与问题最相近的答案。问题和答案的表示使用词袋法 (Bag Of Words，BOW), 使用成对的训练 (pairwise training) 方法，即一个正例随机选取多个知识库中的事实反例进行训练。Bordes[9] 等人意识到仅仅使用自身的答案三元组表示答案向量过于简单，因此引入答案结点的子图，选取与结点距离一跳 1-hop[9]和两跳 2-hop[10]的结点，将子图结点的关系以及结点本身信息都加入到答案结点，从而更综合的表示答案结点，此处向量的表示仍采用 BOW 方法，该结合子图表示的方法也一定程度上提升了问答性能。

注意到 Bordes 等人 BOW 方法表示问题、答案的缺陷，BOW 方法忽略问题中信息的先后顺序，对于复杂问题表达能力不够，并且 Bordes 等人的模型没有考虑对问句类型进行分析。Dong[10] 等人试图通过从问题的不同方面来表示问题、答案向量，他们考虑从三个方面理解问题，即问题的答案路径（answer path）、问题的答案上下文（answer context）和问题的答案类型（answer type）。问题的答案路径指答案结点和问题主题实体之间的关系集合；问题的答案上下文指与答案路径直接相连的实体集合和关系集合；问题的答案类型指答案的数据类型或者结点类别，如答案为时间类型或类别人等。Dong 等人使用三个不同参数的 CNNs（Convolutional Neural Networks）分别表示问题的这三个方面向量，同时也表示答案的这三个方面向量，最后用问题和答案这三个方面对应向量的内积操作和表示问题和答案的相似度，用来选择最为相似的候选答案。

上述方法都是在试图根据问题或者答案的相关方面来更加综合的表示问题或者答案。从实质上来说，问题的表示和答案的表示都是单独进行，也就是说问题的表示没有参照当前的候选答案信息，候选答案的表示也没有参照当前的问题信息。

随着深度学习技术的进一步发展，在神经机器翻译领域（neural machine translation,

[7]https://developer.google.com/freebase

[8]http://nlp.stanford.edu/

[9]直接与结点相连的结点，即与结点路径长度为 1 的结点

[10]与结点路径长度为 2 的结点

NMT），一种注意力（Attention）机制被证明在机器翻译任务上面有不错的性能提升。如 Bahdanau[43] 等人将 Attention 运用到传统的基于编码-解码（encoder-decoder）簇的机器翻译中，传统的解码器在生成翻译词的时候是将源句子表示成固定长度向量，该研究猜想将源句子表示成固定长度可能是性能提升的瓶颈，因此提出了注意力机制，在预测翻译目标词的时候，通过注意力模型自动地搜索与目标词相关的源句子中的部分词。在英语到法语的翻译任务中，该研究取得了较好的性能。

Luong[44] 等人更进一步研究了两种类别的注意力机制（全局注意力、局部注意力）在机器翻译任务上的效果。该研究使用全局注意力每次都关注整个源句子词，使用局部注意力每次只关注源句子中的部分词，该研究也证实了这两种注意力机制对英语到德语的翻译均有效，最终该研究使用两种注意力的集成模型，在 WMT'15[11]英语到德语的翻译任务上，取得了较好的性能效果。

除了机器翻译任务，注意力机制也被运用到句子级别的摘要（sentence-level summarization）任务上, 并取得了一定的性能提升。Rush[45] 等人将局部注意力机制运用到句子摘要任务中，在生成每个摘要词的时候对齐源句子中的部分关键词，在 DUC-2004[12]任务上, 也取得了高于 baseline 的性能提升。

受到注意力机制的启发，有研究将注意力机制运用到知识库问答中，通过注意力机制动态的根据答案向量表示问题向量或者根据问题向量表示答案向量，避免了之前工作中独立表示问题向量和答案向量的缺陷。liu[46] 等人根据问题每个词对答案的注意力分配来综合表示问题。该研究首先将答案向量表示成四部分，即答案实体（answer entity）、答案关系（answer relation）、答案类型（answer type）和答案上下文（answer context，该论文中为与答案实体直接相连的实体集合），然后问题中每个词的表示根据答案四个方面的综合注意力权重求和，最后使用问题相对答案四个方面的表示向量与答案的四个方面向量求内积和得到问题、答案的相似度，并根据设定的答案相似度边距（margin）来确定最终答案。该文献中为解决未登陆词 (out of vocabulary，OOV) 问题，加入全局的知识库作为训练的知识源。该研究在 freebase 知识库问题任务中取得了同类基于信息检索的方法中较好的性能。Tan[47] 等人将注意力机制运用到深度学习模型中回答非事实型问题。该研究先是通过基本框架双向循环神经网络模型来表示问题、答案的分布式词向量，然后在答案的表示时，使用注意力机制来根据问题的内容生成答案向量，最后根据问题、答案的余弦值计算两者的相似度。在 TREC-QA[13]和 InsuranceQA[14]任务上，此研究模型的实验效果超过了一些 baseline 模型。

通过以上对基于语义解析的 KBQA 和基于信息检索的 KBQA 的方法分析可知，虽然基于语义解析的方法具有一定的有效性，但它们很大程度上受限于外部资源的质量，而且需要大量的特征工程工作，再者引入大量语言学知识会大大增加系统复杂度。而基

[11] http://www.statmt.org/wmt15/translation-task.html

[12] https://duc.nist.gov/duc2004/tasks.html

[13] https://trec.nist.gov/data/qa.html

[14] https://github.com/shuzi/insuranceQA

于信息检索的方法更简单，不需要使用大量的人工特征和语言学特征，实现也更灵活，在开放域知识库 Freebase 上的问答研究实验也表明，该方法可以达到与基于语义解析方法相近的 F 值[10;11]，因此基于信息检索的 KBQA 运用比较广泛。

2.5 本文与已有工作不同

本文两个主要任务为构建地理本体 CGeoOnt、根据构建的地理知识库构建问答系统，本节分别从这两个任务阐述本文与已有工作的不同。

对于构建地理本体 CGeoOnt 任务，本文参照高考地理考题的考察方式组织，一些术语的名字组织也并不是仅仅根据地理教材中的名称，更不是凭主观想象，而是从更方便解题角度组织。再者，地理本体 CGeoOnt 中含有很多特殊元素，如静态属性，专门为类添加的属性描述，类的静态属性性质默认施加到类的实例上。还有地理中的过程表示，地理中的大量现象形成很复杂，很多是带有条件、过程性的，本文针对地理过程也定义了一下特殊的词汇描述，这也是一般的本体构建中没有的。正是因为地理知识组织的复杂性，本文的一些本体元素组织往往超过了一般本体语言的语法定义，本文也针对做了扩充（如添加静态属性），因此本文也不太适合用一般的本体构建工具，如 protege 等，本文需要开发自己独特的本体解析工具。

对于基于地理知识库的问答任务，从本文收集的 web 地理问题分析可知，地理问题表达形式多样，一个地理三元组知识往往可以成为多个问题的答案，如“亚热带季风气候在发展农业生产方面有什么优势”和“温带季风气候在农业生产方面的显著优势是 _ 百度知道”，这两个问题都可以用“季风气候 生产优势 夏季高温多雨，雨热同期。”这个三元组知识作为答案。虽然，两个问题问法不同，但相对答案而言它们等效。这也说明，在我们做问答时，单独的表示问题和答案向量是不充分的，至少是不能表示问题和答案之间的关系。因此，本文在表示地理问题、答案时，充分考虑到了两者的内在关联。再者，本文地理问题本身表达方式偏口语化严重，其更具挑战，且问题中无关信息很多，对答案和问题的表示要求更高。

最后，目前中文地理问答数据集缺乏，为了验证实验的有效性，本文还需构建客观、真实、多样化的 web 地理问题、答案对数据集。

2.6 本章小结

本章介绍了基于本体的地理知识问答的相关研究现状。具体包括本体构建的研究现状和知识库问答的研究现状。本章首先介绍本体的四个基本元素个体、类、属性、关系和常见的本体描述语言；然后介绍本体构建的主流方法，主要是介绍八种常见的人工本体构建方法的优缺点和适合范围；再然后介绍知识库问答的两个主流方法——基于语义解析的 KBQA 问答和基于信息检索的 KBQA 的操作流程以及各方向的研究方法优缺点；最后介绍本文工作与已有工作的不同之处。

第三章　基于本体的地理知识问答

基于本体的地理知识问答指的是：根据构建的地理本体知识库，构建一个基于知识库的问答系统，针对用户提出的地理问题给出精准的答案。本章首先介绍本文的具体任务和系统结构图，然后介绍地理本体知识库的具体构建方法，最后介绍基于地理本体知识库问答系统的实现流程。

3.1　论文任务

本论文任务是：根据用户提出的地理问题，系统直接给出该问题的简短、精准答案。如下为用户所提出的四个问题和系统相应回复答案举例：

（1）提问：“地球的半径大约是多少呢?”，回复：“6371km”

（2）提问：“地球半径约多少米多少千米?”，回复：“6371km”

（3）提问：“怎样保护热带雨林?”，回复：“加强环境教育，提高公民环保意识; 加强雨林管理和保护，建立自然保护区。”

（4）提问：“中国采取了哪些措施保护热带雨林?”，回复：“加强环境教育，提高公民环保意识; 加强雨林管理和保护，建立自然保护区。”

问题（1）、（2）问的为同一个问题，均可根据本文知识库中三元组“（地球，半径，6371km）”来回答。问题（3）、（4）也是问的相同问题、只是语言不同，均可根据知识库中三元组“（雨林，保护措施内容，“加强环境教育，提高公民环保意识; 加强雨林管理和保护，建立自然保护区”）”来回答，系统需要区分知识库中“雨林”和问题中问的“热带雨林”其实为同一个概念。因此，本文主要有两个任务，如下：

（1）构建可以回答地理问题的地理本体知识库。

（2）在构建的地理本体知识库上构建问答系统，系统需满足用户按不同语言组织方式随意提问，均能正确匹配知识库中可以回答问题的答案三元组。

3.2　系统结构图

针对基于本体的地理知识问答的两个任务，本文设计了如下图3.1所示的系统结构图。该系统结构图由两个模块组成，地理本体知识库构建模块、基于注意力机制的地理知识库问答模块。第二个模块包括两个子模块：地理问题输入、处理模块和答案查询模块

地理本体知识库构建模块包括三个任务，第一是根据高中地理相关知识源人工构建地理本体 CGeoOnt，第二是将本体 CGeoOnt 和基于百度百科自动构建的地理本体 Clinga

融合成更大的本体并进行本体存储和检索，第三个任务是根据融合而成的地理本体构建本体词汇表和本体同义词词汇表，本文统称为为本体词典。

基于注意力机制的地理知识库问答模块的子模块——地理问题输入、处理模块包括两个任务，第一是根据地理核心知识考点，获取跟知识考点相关的真实地理问题集，第二是对问题进行分析，包括问题分词、命名实体识别、词性标注和问题中包含的本体主题词识别，最终目的是生成候选答案实体供候选答案检索。

基于注意力机制的地理知识库问答模块的子模块——答案查询模块包括三个任务，第一是根据候选答案实体去地理本体知识库查询相应的候选答案三元组，第二是构建地理知识库问答模型对用户输入的问题和其候选答案进行向量表示，第三是根据问题和候选答案的向量表示，由相似度打分策略和最终答案选择策略选取最终答案。

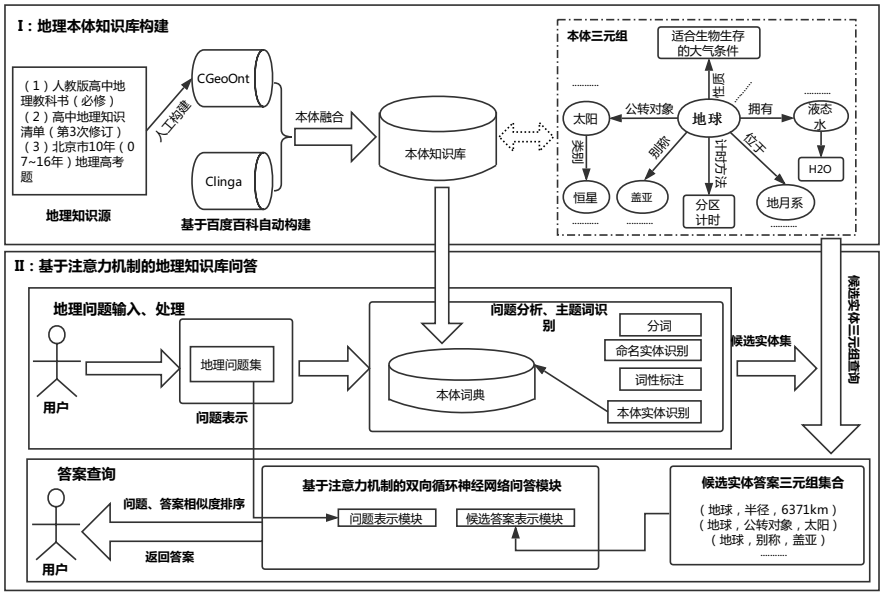


图 3.1: 基于本体的地理知识问答系统结构图

3.3 地理本体知识库构建

地理本体知识库构建内容分四个部分来介绍，先介绍地理本体 CGeoOnt(chinese geographic ontology) 的构建，然后介绍本体 Clinga(chinese linked geographical dataset) 与本体 CGeoOnt 融合，再然后介绍本体的存储与检索，最后介绍本体词典生成方法。

3.3.1 地理本体 CGeoOnt 构建

本文本体构建方法借鉴了斯坦福七步法的流程，但是与斯坦福的七步法有一些区别，如本文没有可以复用的本体，并且七步法需一次性列出本体的重要术语，其可操作

性不强。因此，根据地理领域的特点，本文主要分四个步骤进行本体构建工作, 分别如下：

（1）地理知识源选取。选取本体构建所需要的资料，文本或者图表形式。

（2）地理知识体系定义。确定本体构建的三元组知识组织顺序。

（3）地理本体构建规范定义。约定本体构建的知识组织规范。

（4）地理本体基本元素定义。定义类、属性、关系，创建实例。

3.3.1.1 地理知识源选取

地理本体构建知识源来自教育部高中地理教材以及部分的高考地理试题。地理教材包括人教版高中地理必修一自然地理、必修二人文地理、必修三区域可持续发展、区域地理、选修三旅游地理、选修五自然灾害与防治、选修六环境保护这七本教材和一本高中地理知识清单。高中地理知识清单是专家对上述七本高中地理教材中核心考点的精确提炼，以及对每本书知识点的解题方法讲解。高考试题选取的是北京市近 10 年（2007–2016 年）的地理高考题，选取高考题作为知识源的目的是：标注人员可以根据高考题快速把握高考地理考点以及弄清考点所需要的辅助知识，因此能够有针对性地去教材中寻找需要标注的核心知识，避免标注大量对于解题无用的地理知识，大大提高了地理核心知识库构建的效率。本文地理资源均为电子资源，且格式为 HTML 的网页形式，里面包含文字叙述、图和表格。

3.3.1.2 地理知识体系定义

地理知识体系的定义根据《高中地理知识清单（第 3 次修订）》中的组织结构进行，该高中地理知识清单包括了整个高中地理教科书中知识的组织顺序，如高中地理教材的组织先从必修地理开始介绍，包括必修自然地理、必修人文地理、必修区域可持续发展。然后是介绍选修地理部分，包括选修旅游地理、选修自然灾害与防治和选修环境保护。对于每一本教材中的知识体系根据书中每个章节的相关主题、相关知识点和相关方法三个方面来组织，这样的好处是使知识库的知识都能够找到其知识的出处，便于后续分析知识之间的关系和知识溯源，保留知识的完整性。如在标注概念类“天体”的知识三元组时，需要标注出“天体”在书中所属的“相关主题”是“宇宙中的地球”，以及“天体”涉及的“相关知识点”是“地球的宇宙环境”和跟“天体”相关联的“相关方法”是“天体类型的判断方法”。

3.3.1.3 地理本体构建规范

地理本体构建规范主要为保证最终本体的构建形式一致，因为本体构建需要多个专业人员参与，不同人的标注主观性很大，必须经过统一的规范约束。首先，本文地理知识需要较强的表达能力且还需具备很强的推理能力，因此本文约定本体描述语言使用

OWL，构建符合 RDFS 和 OWL DL 级别规范。其次，由于本文地理本体构建有较多本领域的特殊定义，有些定义需要突破 OWL DL 的规范，因此本文构建工作不使用当前存在的本体构建工具，如 protege 等，本文直接以 RDF 三元组的语法形式 Turtle[1] 来表示每条本体知识三元组。最后，本文规定个体、类、属性等的命名说明规范，统一本体的最终表现形式。

为简化三元组的表示，本文定义了本文地理本体术语的命名空间，包括自定义的三个地理本体元素——个体、属性及类的命名空间 gsr、gss、gso，以及包括七个常见的本体元素命名空间：rdf、rdfs、owl、skos、xs、op 和 fn。以上命名空间如下所示：

@prefix gsr: <http://ws.nju.edu.cn/geoscholar/resource>.
@prefix gss: <http://ws.nju.edu.cn/geoscholar/staticOntology>.
@prefix gso: <http://ws.nju.edu.cn/geoscholar/ontology>.
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns>.
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema>.
@prefix owl: <http://www.w3.org/2002/07/owl>.
@prefix skos: <http://www.w3.org/2004/02/skos/core>.
@prefix xs: <http://www.w3.org/2001/XMLSchema>.
@prefix op: <http://www.w3.org/2002/08/xquery-operators>.
@prefix fn: <http://www.w3.org/2005/xpath-functions>.

以下为其他约定，核心思想为使本体不同元素有区分度，同时见其英文名可知其中文义：

- 个体：以“R_”开头，如 gsr:R_ 太阳。
- 类：以“C_”开头，如 gso:C_ 天体。
- 属性（静态属性）：以“P_”开头，两个属性 owl:ObjectProperty、owl:DatatypeProperty 分别以“P_o_”和“P_d_”开头。如 gso:P_o_ 公转对象，gso:P_d_ 高度。
- 个体须声明类别 owl:Class、属性必须确定是对象属性 owl:ObjectProperty 还是数据类型属性 owl:DatatypeProperty。
- 每个术语有且仅有一个 skos:prefLabel，但是可以有多个 skos:altLabel 用于别名。
- 每个术语有必要时使用 skos:definition 给出其定义，使用 rdfs:comment 对其重要方面进行评论。
- 术语间多使用 gso:relatedTo 来表达它们有一定的关联性。
- 为避免处理的复杂性，三元组中不使用空白结点（blank node ）。

[1] https://www.w3.org/TR/turtle/

3.3.1.4 地理本体基本元素定义

本文地理本体构建遵循自下而上的构建原则，按照书本知识的组织顺序构建本体，不一次性的定义其领域所有术语，而采取根据每章、每节中的考点知识构建，考点知识的确定参照近 10 年的北京市地理高考题以及书中重点强调的考点。下面分别介绍类、属性、个体和本文特殊地理元素的具体构建。

（1）类的构建

类的构建需要声明其类型是 OWL 中的类,并且需要定义其唯一的名称 skos:prefLabel，而且该类若有别名则需定义别名 skos:altLabel，该类的父类若存在则需要定义出。类相关的定义或者评论可以通过 skos:definition 和 skos:comment 定义出，若该类所对应的课文章节主题和知识点为重要考点，也应该定义出。如下举例为本文对概念类“行星”的三元组知识表示，知识三元组中的主语、谓语、宾语之间以空格表示：

gso:C_ 行星　rdf:type　owl:Class　.

gso:C_ 行星　skos:prefLabel　”行星”^^xs:string　.

gso:C_ 行星　rdfs:subClassOf　gso:C_ 天体　.

gso:C_ 行星　skos:definition　”在椭圆轨道上，环绕恒星运行、近似球状的天体，其质量比恒呈小，本身是不发光的^^xs:string　.

gso:C_ 行星　gsb:relatedToKPoint　gsb:M_ 太阳对地球的影响　.

gso:C_ 行星　gsb:relatedToKPoint　gsb:M_ 宇宙中的地球　.

gso:C_ 行星　gsb:relatedToTopic　gsb:M_ 行星地球.

（2）属性的构建

本文属性包括普通的属性和本文特殊定义的属性，此节只讲述普通属性的构建，特殊定义的属性见第四小节“地理特殊定义元素”介绍。本文普通属性的构建需明确区分是对象属性 owl:ObjectProperty 还是数据类型属性 owl:DatatypeProperty，属性的性质（如对称性、传递性等）需要定义出来，便于根据属性进行相关推理。同时，对于对象属性，本文需定义属性的定义域 rdfs:domain 和值域 rdfs:range，不好区分情形下使用总类 owl:Thing 描述；对于数据类型属性，数据类型的单位必须定义，如下分别为对象属性“垂直”和数据属性“宽度值”的举例表示：

对象属性：垂直

gso:P_o_ 垂直　rdf:type　owl:ObjectProperty　.（对象属性）

gso:P_o_ 垂直　skos:prefLabel　”垂直”^^xs:string　.（属性名称）

gso:P_o_ 垂直　rdfs:domain　owl:Thing　.　（定义域为所有对象）

gso:P_o_ 垂直　rdfs:range　owl:Thing　.　（值域为所有对象）

gso:P_o_ 垂直　a　owl:SymmetricProperty　.　　（对称性）

gsr:R_ 水平气压梯度力　gso:P_o_ 垂直　gsr:R_ 等压线　.（属性连接的两个对象）

gso:P_o_ 垂直　gsb:relatedToKPoint　gsb:M_ 大气的水平运动-风　.（属性涉及到的

考点知识）

数据属性：宽度值

gso:P_d_ 宽度值　rdf:type　owl:DatatypeProperty　.（数据属性）

gso:P_d_ 宽度值　skos:prefLabel　”宽度值”^^xs:string　.（属性名称）

gso:P_d_ 宽度值　gso:physicalQuantity　”宽度”^^xs:string　.（属性单位名称）

gso:P_d_ 宽度值　gso:unit　”米”^^xs:string　.（属性单位类型）

gso:P_d_ 宽度值　rdfs:domain　owl:Thing　.（定义域为所有对象）

gso:P_d_ 宽度值　rdfs:range　xs:double　.（值域为双精度数值）

gso:P_d_ 宽度值　rdfs:comment　”描述某物 a 的宽度为某个值”^^xs:string　.（属性的一般性评论）

gsr:R_ 南极附近海上最大的冰山　gso:P_d_ 宽度值　”97000.0”^^xs:double　.（属性描述的个体知识）

gso:P_d_ 宽度值　gsb:relatedToKPoint　gsb:M_ 水资源与人类社会　.（属性相关的知识考点）

gso:P_d_ 宽度值　gsb:relatedToTopic　gsb:M_ 水资源的合理利用　.（属性相关的知识考点章节）

（3）个体的构建

个体的构建只包含标注章节中包含高考地理考点的专业术语，其它非考点术语忽略不予标注。本文个体构建遵循的重要原则是每个个体须申明其类别，而且其类别允许为多个，这一点也符合现实世界的表示。如下为个体“地球”的知识表示：

gsr:R_ 地球　skos:prefLabel　”地球”^^xs:string　.　（个体名称）

gsr:R_ 地球　rdf:type　gso:C_ 行星　.（个体的类别，此处地球有三个标签类别）

gsr:R_ 地球　rdf:type　gso:C_ 物体　.（个体的类别）

gsr:R_ 地球　rdf:type　gso:Location　.（个体的类别）

gsr:R_ 地球　gso:P_d_ 性质　”适合生物生存的大气条件”^^xs:string　.（个体数据-字符串属性, 描述的事实是：地球拥有适合生物生存的大气条件的性质）

gsr:R_ 地球　gso:P_d_ 轨道偏心率　”0”^^xs:string　.（个体数据-字符串属性）

gsr:R_ 地球　gso:P_d_ 计时方法　”分区计时”^^xs:string　.（个体数据-字符串属性）

gsr:R_ 地球　gso:P_d_ 时区数量　”24”^^xs:integer　.（个体数据-整型属性）

gsr:R_ 地球　gso:P_d_ 平均密度　”5”^^xs:string　.（个体数据-字符串属性）

gsr:R_ 地球　gso:P_o_ 公转对象　gsr:R_ 太阳　.（个体的对象属性）

gsr:R_ 地球　gso:P_o_ 天然卫星　gsr:R_ 月球　.（个体的对象属性）

gsr:R_ 地球　gso:P_o_ 拥有　gsr:R_ 液态水　.（个体的对象属性）

gsr:R_ 地球　gso:P_o_ 南部　gsr:R_ 南半球　.（个体的对象属性）

（4）地理本体特色元素

□ 静态属性

静态属性，本文前缀表示为 gss，是为增强地理知识的表达能力而设定的，需与普通属性 gso 相区别。静态属性严格上并不满足 OWL 的属性规范，因为 OWL 的属性连接的可以是两个个体或者一个个体和一个数据类型数据，而本文的静态属性描述的是类的属性特征。静态属性重要的特性是：静态属性所描述的类所包含的所有个体均具有该静态属性特征，这一点类似于个体继承类的特征。由于静态属性是对一类个体的描述，本文可以允许其不申明定义域 rdfs:domain 和值域 rdfs:range，但是该静态属性仍然区分是对象属性 owl:ObjectProperty 还是数据类型属性 owl:DatatypeProperty，如下对类“暖流”定义了一个静态属性“作用”。

gss:P_d_ 作用　rdf:type　owl:DatatypeProperty　.（gss 申明静态数据类型属性）

gso:C_ 暖流　gss:P_d_ 作用　”对沿岸气候增温增湿”^^xs:string　.

gsr:R_ 北大西洋暖流　rdf:type　gso:C_ 暖流.（申明暖流的个体-北大西洋暖流）

则表明：北大西洋暖流同样具有暖流的静态属性特性，也即：

gsr:R_ 北大西洋暖流　gso:P_d_ 作用　”对沿岸气候增温增湿”^^xs:string　.

□ 封闭集

封闭集用符号 gso:isClosed 表示，包括类和属性的封闭集。类的封闭集表示某个类的实例可以全部列举出，属性的封闭集表示某个个体在该属性上取值是封闭的，即包含有限的取值且可以穷举出来。

● 类封闭集知识举例：地球的极点包括南极点、北极点

gsr:R_ 北极点　a　gso:C_ 地球极点　.

gsr:R_ 南极点　a　gso:C_ 地球极点　.

gso:C_ 地球极点　gso:isClosed　”true”^^xs:boolean　.

地球极点类在封闭集 gso:isClosed 属性上取值为 true，说明地球极点的取值实例个体已经完整了，只有南极点和北极点。

● 属性封闭集知识举例：太阳大气层由里到外可分为光球、色球、日冕三层。

gsr:R_ 太阳大气层　gso:P_o_ 包含　gsr:R_ 光球　;

gsr:R_ 太阳大气层　gso:P_o_ 包含　gsr:R_ 色球　;

gsr:R_ 太阳大气层　gso:P_o_ 包含　gsr:R_ 日冕　.

gso:P_o_ 包含　gso:isClosedAt　gsr:R_ 太阳大气层　.

该知识表示“包含”属性在个体“太阳大气层”上处于封闭状态，说明太阳大气层只包括光球、色球和日冕三层。

□ 地理数值类型

常见的数值类型属性的取值都较为明确，如使用整型 xs:integer，双精度浮点型 xs:double 来表示。然而，在地理学科中，大多数出现的数值都不是精确值，往往是某种程度上的近似值。下面介绍本文几种类型的近似值处理介绍：

● 约取值

约取值的主要描述特征为“某某约为多少”，此种情况本文视作精确值处理，如知识：太阳表面的温度大约为 6000K，直接取 6000K 作为温度数值, 表示如下：

gsr:R_ 太阳　gso:P_d_ 表面温度　”6000”^^xs:decimal .

● 表示“几”

对于地理知识源中包含“几”的数值描述，本文使用两种方法进行标注：第一种是查阅相关的权威资料，明确其具体的值；第二种是在无法通过其它辅助资料查询到相关精确数值时，使用范围 1—9 加上单位的范围值表示。如下两种举例：

地理知识：色球厚度约为几千千米，实际上通过查阅维基百科可以发现，“色球厚度大约是 2,000 公里”，因此可以表示如下：

gsr:R_ 色球　gso:P_d_ 厚度　“2e6”^^xs:double　.

如果无法查阅其他资料得到准确值，则采取范围区间值来表示，如下：

gsr:R_ 色球　gso:P_d_ 厚度　”(1e6, 9e6)”^^ xs:impreciseInterval　.

● 范围值

本文标注过程中主要遇到三种范围值的表述，第一种为取值唯一，但不精确，第二种为条件缺省时导致的取值不确定，第三种为无法精确到确定的取值区间。

第一种范围值使用不精确的区间值 xs:impreciseInterval 表示，以圆周率举例，圆周率的值为 3.1415926—3.1415927 范围之间。或者如地理知识：崇明岛全岛的面积为 1200 多平方千米，可以表示如下：

gsr:R_ 崇明岛　gso:P_d_ 面积值　”(1.201e9, 1.299e9)”^^xs:impreciseInterval　.

第二种范围值使用不明确的区间值 xs:unclearInterval 表示，如日地距离的取值，（时间不同时，日地距离亦不同），某省某刻的温度（具体的地点不同时，温度也不同）。这样的例子当补充了具体的条件，如具体时间、具体地点后就有可能转换为精确值。本文把此种复杂的条件限制使用最低到最高的区间值来表示，如下表示地球大气上界的高度在不同时间的范围值：

gsr:R_ 地球大气上界　gso:P_d_ 高度　”(2e6, 3e6)”^^xs:unclearInterval　.

第三种使用不能被精确为确定值的区间 xs:scopeInterval 表示，如没有最大值的开区间，以地理知识为例：亚热带季风气候年降水量在 800mm 以上，范围值只有下界值，没有上界值，此种形式本文使用开区间表示，如下：

gsr:R_ 亚热带季风气候　gso:P_d_ 年降水量　”[0.8,)”^^xs:scopeInterval　.

● 地理区域表示

地理区域主要指存在于空间中且可以准确定位的空间位置。地理中常常考察地理地点之间的位置关系（接壤、方向、包含等），因此对于地理重要地点的位置表示，需要准确细致的刻画。本文将空间所有事物定义为一个大类，记作 gso:SpatialThing，将可以空间定位的地点定义为空间事物类的子类，记作 gso:Location。如下先介绍空间位置关系，后介绍利用空间关系来表示空间位置。

空间关系：

空间关系用于描述两个可定位地点位置的相对方位关系，从某种程度上来看，它也是一种对象属性，表达两个空间位置的关系。本文定义的位置关系都符合地理领域的方位表达方式，本文以常见的方位关系“位于”、“中部”来举例说明。“位于”表示一个位置在另一个位置上，属于一种包含关系。“中部”表示对一个位置的方向约束，如中国中部等。如下为具体的地理知识表示：“基拉韦厄火山位于北太平洋中部的夏威夷群岛上”，本文需要准确表示出“位于”关系，北太平洋中部和被太平洋的方位“中部”关系

gsr:R_ 基拉韦厄火山　rdf:type　gso:Location　.

gsr:R_ 夏威夷群岛　rdf:type　gso:Location　.

gsr:R_ 北太平洋　rdf:type　gso:Location　.

gsr:R_ 北太平洋中部　rdf:type　gso:SpatialThing　.

gsr:R_ 基拉韦厄火山　gso:P_o_ 位于　gsr:R_ 夏威夷群岛　.

gsr:R_ 夏威夷群岛　gso:P_o_ 位于　gsr:R_ 北太平洋中部　.

gsr:R_ 北太平洋　gso:P_o_ 中部　gsr:R_ 北太平洋中部　.

空间位置：

对于空间位置，尤其是需要通过复杂空间关系来表示的空间位置，先得将其声明为 gso:SpatialThing 的实例，然后通过空间关系属性（谓词）建立与其它空间位置的关系，最后通过与其它位置构成的关系来整体表示该空间位置。如地理知识：“亚热带季风气候分布地区为我国秦岭-淮河以南、朝鲜半岛南部、日本群岛南部”，该知识中包含的“秦岭-淮河以南”、“朝鲜半岛南部”和“日本群岛南部”这三个带有方向限定的位置，需以本文定义的空间关系“南方”、“南部”并结合具体位置来联合表示。比如“秦岭-淮河以南”需要定义为“秦岭-淮河”的空间关系“南方”，具体表示如下：

gsr:R_ 秦岭-淮河　rdf:type　gso:Location　.

gsr:R_ 日本群岛　rdf:type　gso:Location　.

gsr:R_ 朝鲜半岛　rdf:type　gso:Location　.

gsr:R_ 秦岭-淮河以南　rdf:type　gso:SpatialThing　.

gsr:R_ 日本群岛南部　rdf:type　gso:SpatialThing　.

gsr:R_ 朝鲜半岛南部　rdf:type　gso:SpatialThing　.

gsr:R_ 秦岭-淮河 gso:P_o_ 南方 gsr:R_ 秦岭-淮河以南 .（空间关系“南方”和空间位置“秦岭-淮河”联合表示出空间位置“秦岭-淮河以南”）

gsr:R_ 日本群岛 gso:P_o_ 南部 gsr:R_ 日本群岛南部 .

gsr:R_ 朝鲜半岛 gso:P_o_ 南部 gsr:R_ 朝鲜半岛南部 .

gsr:R_ 亚热带季风气候 gso:P_o_ 分布地区 gsr:R_ 秦岭-淮河以南 ,

gsr:R_ 亚热带季风气候 gso:P_o_ 分布地区 gsr:R_ 日本群岛南部 ,

gsr:R_ 亚热带季风气候 gso:P_o_ 分布地区 gsr:R_ 朝鲜半岛南部 .

再如“接壤”位置之间的表示，如知识：“亚欧板块和太平洋板块的交界处地震多发”，需要表现“亚欧板块”与“太平洋板块”的” 相接” 也即“接壤”关系。

gsr:R_ 亚欧板块 rdf:type gso:Location .

gsr:R_ 太平洋板块 rdf:type gso:Location .

gsr:R_ 亚欧板块 gso:P_o_ 相接 gsr:R_ 太平洋板块 .

gsr:R_ 亚欧板块和太平洋板块的交界处 rdf:type gso:SpatialThing ;

gsr:R_ 亚欧板块和太平洋板块交界处 gso:P_o_ 相接 gsr:R_ 亚欧板块 ;

gsr:R_ 亚欧板块和太平洋板块交界处 gso:P_o_ 相接 gsr:R_ 太平洋板块 ;

gsr:R_ 亚欧板块和太平洋板块交界处 gso:P_d_ 地震发生率 ” 多”^^xs:string .

● 过程表示

本文过程是指地理中一些现象在一系列条件下通过一系列步骤而形成的总称，常常有顺序过程和循环过程。本文将过程定义为一个大类，记作 gso:Process，顺序过程记作 gso:SequentialProcess，循环过程记作 gso:CircularProcess，分别声明为过程类的两个子类。对于顺序过程或者循环过程，都需要严格的定义其形成所需要的步骤，以及每个步骤上面的步骤条件。具体定义如下：

先声明类:

gso:Process rdf:type owl:Class . （过程类）

gso:CircularProcess rdfs:subClassOf gso:Process .（循环型过程类）

gso:SequentialProcess rdfs:subClassOf gso:Process . （顺序型过程类）

gso:StepOfProcess rdf:type owl:Class .（过程的步骤类）

再声明条件、步骤属性:

gso:condition4Step rdf:type owl:DatatypeProperty . （步骤条件，取值为字符串）

gso:condition4Step rdfs:domain gso:StepOfProcess .

gso:stepDesc rdf:type owl:DatatypeProperty . （步骤文字描述，取值为字符串）

gso:stepDesc rdfs:domain gso:StepOfProcess .

gso:stepNum rdf:type owl:DatatypeProperty . （步骤的序号，第几个步骤，取值为正整数）

gso:stepNum rdfs:domain gso:StepOfProcess .

如下复杂地理过程知识的具体表示：“冷锋过境前，会受单一暖气团控制，温暖晴朗。当冷气团主动移向暖气团时，较重的冷气团会插入暖气团下面，使暖气团被迫的抬升。暖气团在抬升的过程中会逐渐的冷却，其中水汽易凝结成云。如果暖空气中含有大量水汽，那么可能会导致雨雪天气。冷锋移动速度较快，常常带来较强的风。冷锋过境后，冷气团替代原来暖气团的位置，气压升高，气温降低，湿度骤降，天气则转好。”

上述叙述的是地理现象“冷锋过境”的形成过程，如下具体表示：

先声明“过程”和“步骤”：

gsr:R_ 冷锋过境过程　rdf:type　gso:SequentialProcess　.

gsr:R_ 冷锋过境过程　gso:numberOfSteps　”3”^^xs:positiveInteger　.

gsr:R_ 冷锋过境步骤 _1　rdf:type　gso:StepOfProcess　.

gsr:R_ 冷锋过境步骤 _2　rdf:type　gso:StepOfProcess　.

gsr:R_ 冷锋过境步骤 _3　rdf:type　gso:StepOfProcess　.

gsr:R_ 冷锋过境过程　gso:includesStep　gsr:R_ 冷锋过境步骤 _1　,

gsr:R_ 冷锋过境过程　gso:includesStep　gsr:R_ 冷锋过境步骤 _2　,

gsr:R_ 冷锋过境过程　gso:includesStep　gsr:R_ 冷锋过境步骤 _3　.

再申明“步骤”的文字描述 gso:stepDesc 和条件 gso:condition4Step：

gsr:R_ 冷锋过境步骤 _1　gso:stepNum　”1”^^xs:positiveInteger　;

gsr:R_ 冷锋过境步骤 _1　gso:stepDesc　”当冷气团主动移向暖气团时，较重的冷气团插入暖气团下面，使暖气团被迫抬升。”^^xs:string　.

gsr:R_ 冷锋过境步骤 _2　gso:stepNum　”2”^^xs:positiveInteger　;

gsr:R_ 冷锋过境步骤 _2　gso:stepDesc　”暖气团在抬升过程中逐渐冷却，其中水汽容易凝结成云。”^^xs:string　.

gsr:R_ 冷锋过境步骤 _3　gso:stepNum　”3”^^xs:positiveInteger　;

gsr:R_ 冷锋过境步骤 _3　gso:condition4Step　”暖空气含大量水汽。”^^xs:string　;

gsr:R_ 冷锋过境步骤 _3　gso:stepDesc　”如果暖空气中含有大量的水汽，那么可能会带来雨雪天气。”^^xs:string　.

此“冷锋过境”过程中实际上也包含了不同时段的“天气变化”，因此“冷锋过境中的天气变化”也需要表示为一个顺序过程，这样在考察“冷锋过境”与天气变化的关系是，可以运用此结构化的知识表示。具体表示如下：

申明“冷锋过境中的天气变化”为顺序过程:

gsr:R_ 冷锋过境中的天气变化　rdf:type　gso:SequentialProcess　;

gsr:R_ 冷锋过境中的天气变化　gso:numberOfSteps　”3”^^xs:positiveInteger　.

gsr:R_ 冷锋过境中天气变化的步骤 _1　rdf:type　gso:StepOfProcess　.

gsr:R_ 冷锋过境中天气变化的步骤 _2　rdf:type　gso:StepOfProcess　.

gsr:R_ 冷锋过境中天气变化的步骤 _3　rdf:type　gso:StepOfProcess　.

gsr:R_ 冷锋过境过程中的天气变化　gso:includesStep　gsr:R_ 冷锋过境中天气变化

的步骤 _1 ，

gsr:R_ 冷锋过境过程中的天气变化 gso:includesStep gsr:R_ 冷锋过境中天气变化的步骤 _2 ，

gsr:R_ 冷锋过境过程中的天气变化 gso:includesStep gsr:R_ 冷锋过境中天气变化的步骤 _3 .

申明“冷锋过境中的天气变化”的每个步骤的内容和条件:

gsr:R_ 冷锋过境中天气变化的步骤 _1 gso:stepNum ”1”^^xs:positiveInteger ;

gsr:R_ 冷锋过境中天气变化的步骤 _1 gso:condition4Step ”过境前”^^xs:string ;

gsr:R_ 冷锋过境中天气变化的步骤 _1 gso:stepDesc ”过境前，单一暖气团控制，温暖晴朗。”^^xs:string .

gsr:R_ 冷锋过境中天气变化的步骤 _2 gso:stepNum ”2”^^xs:positiveInteger ;

gsr:R_ 冷锋过境中天气变化的步骤 _2 gso:condition4Step ”过境时”^^xs:string ;

gsr:R_ 冷锋过境中天气变化的步骤 _2 gso:stepDesc ”过境时，常有阴天、下雨、刮风、降温等天气现象。”^^xs:string .

gsr:R_ 冷锋过境中天气变化的步骤 _3 gso:stepNum ”3”^^xs:positiveInteger ;

gsr:R_ 冷锋过境中天气变化的步骤 _3 gso:condition4Step ”过境后”^^xs:string ;

gsr:R_ 冷锋过境中天气变化的步骤 _3 gso:stepDesc ”过境后，气强和湿度骤降，天气转晴。”^^xs:string .

3.3.2 地理本体融合

本文地理核心知识库由构建的两个本体融合而成。这两个本体分别是本文手工构建的 CGeoOnt 和 863 项目组基于百度百科自动构建的本体 Clinga。本节先介绍 Clinga 的相关情况，后介绍两个地理本体的融合方法。

3.3.2.1 本体 Clinga

本体 Clinga[2]（Chinese linked geographical dataset）为 863 项目组 Hu[48] 等人根据百度百科自动抽取构建的中文地理链接数据集。Clinga 的本体 Schema 为 Hu 等人人工根据地理领域特点定义，他们将地理分为自然地理和人文地理两个大类，然后在这两个大类下再依次细分为多个小类，最后形成 Clinga 的 schema，具体如类层次摘要图3.2所示。通过定义好的本体 schema 对自动抽取到的百度百科概念做分类，只选取满足定义好的本体 schema 类别的实体，最后将每个百科概念页面结构化成三元组形式，最终得到包含 624,391 个实体概念、130 个类别、73,326,425 个三元组的本体。Clinga 中知识的组织形式为主、谓、宾，即（s, p, o）[3]三元组。例如（A69064644，公转对象，A6768329）表

[2]http://ws.nju.edu.cn/clinga/

[3]s 代表 subject，p 代表 predicate，o 代表 object

示“地球绕着太阳公转”，A6906464 和 A6768329 分别对应地球、太阳在知识库中的存储 ID，“公转对象”表示地球和太阳的一种属性关系。

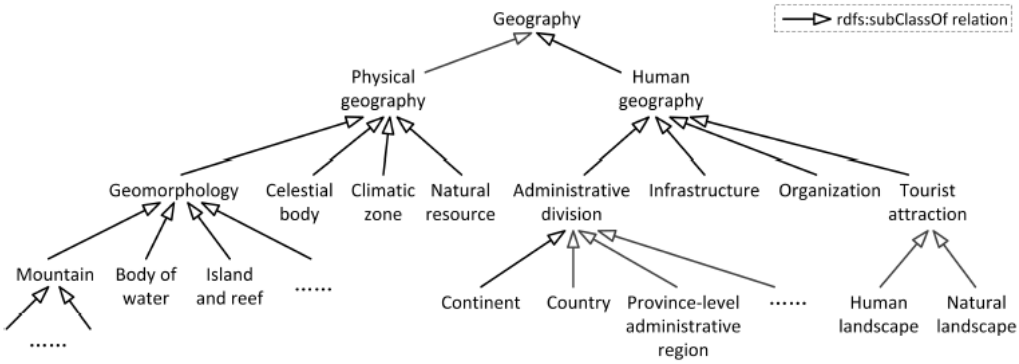


图 3.2: Clinga 本体层次结构定义摘要

图3.2仅仅为 Clinga 本体结构的一个摘要，有向图的根结点为总类别地理，然后箭头被指向的是其两个子类自然地理（physical geography）和人文地理（human geography），然后依次为这两个子类的子类，依此类推。实际图中的类共有 130 个，其中的 35 个叶子类为 GeoNames[4]中没有的类，更具体的情况请参加 Hu[48] 等人的工作。

3.3.2.2 地理本体融合方法

Clinga 与 CGeoOnt 的 Schema 定义存在一定差异，本文需要将这两个本体进行融合得到一个更大的地理本体知识库。本文构建的 CGeoOnt 包含类和属性 3,000 余个，相关 RDF 三元组 25,000 余条，概念类的定义粒度相比 Clinga 而言更细。由于 CGeoOnt 本体相对规模较小，本文采取将本体 CGeoOnt 的 Schema 映射到 Clinga 的 Schema 上，具体操作流程如下：

（1）Clinga 中实体只保留其在百度百科 infobox 域中信息三元组，去掉其 section 域构成的三元组。

（2）运用启发式规则，根据 CGeoOnt 中实体的名称将其类别映射到 Clinga 中的类别。如，CGeoOnt 中“黄山”会被映射到 Clinga 中的类“山”。

（3）人工校验（2）中映射结果。

（4）得到最终本体知识库三元组 4,850,435 条。

3.3.3　地理本体存储与检索

本文采用对象关系数据库 OpenLink Virtuoso[5]存储标注好的地理本体知识三元组，由于 OpenLink Virtuoso 是一个可伸缩的高性能、兼容性强的对象关系数据库引擎，可以提

[4]http://www.geonames.org/

[5]https://virtuoso.openlinksw.com/

供复杂的 SQL、XML、RDF 数据库管理功能，著名的开放域知识库 freebase 也采用此数据库存储，本文运用其存储 RDF 数据的功能，同时还使用 SPARQL[6]语句检索本体知识库中的概念和概念关系三元组信息数据。

3.3.4 地理本体词典生成

本文需要根据构建的地理知识库生成地理概念词典，这些概念包含知识库中的实体和类及两者的同义词。生成的地理本体词典主要用于判断某个命名实体是否属于本文地理知识库的范畴，以便提供相关候选实体的知识库信息，同时也便于查询同一个实体的不同别名。

地理本体词典生成的主要工作为抽取地理概念的同义词，由于本文的知识库有两部分数据来源，分别是本体 Clinga 和本体 CGeoOnt。Clinga 中的概念抽取自百度百科，因此可以通过规则模板来抽取某个概念的同义词，本文统计发现属性名为中文名、英文名、外文名、原名、别名、中文名称、其他名称、别称和又叫，均可以当作概念的同义词。由于百科知识组织的主观性强，有的实体在表示时后面常常跟有括号，如“地球别名盖亚 (Gaia)”，针对此种情况需要将别名值中带有括号的部分也抽取出来，并作为该概念的同义词，如此处的 Gaia 也作为地球的同义词。CGeoOnt 中的同义词已经被标注人员标注出，只需要抽取概念的 skos:altLabel 属性值就可以找出该概念的同义词。通过上述方法，本文得到共包含 75,1103 个概念同义词的本体词典。

3.4 基于注意力机制的地理知识库问答

基于注意力机制的中文地理知识库问答指根据本文构建的中文地理知识库，运用基于注意力机制的双向循环神经网络问答模型，对用户提出的地理问题，从知识库搜索出问题答案三元组的过程。该系统问答流程如图3.3，图3.3表示的是地理问题“2016 年北京处暑节气是什么时间”被回答的流程。首先，需要识别该地理问题中的主题实体，也就是问题是围绕哪个实体概念在提问的，如此问题中考察的是“北京的处暑节气时间”，是考察跟“北京”相关的事实性知识，因此“北京”为主题实体。然后，根据主题实体“北京”从本文构建的知识库中查询出与“北京”直接相连的所有实体作为候选实体，此处为首尔和华盛顿等。最后使用双向循环神经网络对此问题和候选答案三元组——北京、首尔、华盛顿的三元组进行词向量表示，取问题与候选答案三元组相似度最大的三元组作为该问题的答案。

总结本文知识库问答任务，可分为如下四步进行：

（1）识别地理问题中涉及的主题实体，根据主题实体生成候选实体集。

（2）根据候选实体集查询知识库，返回候选实体三元组。

[6]https://www.w3.org/TR/rdf-sparql-query/

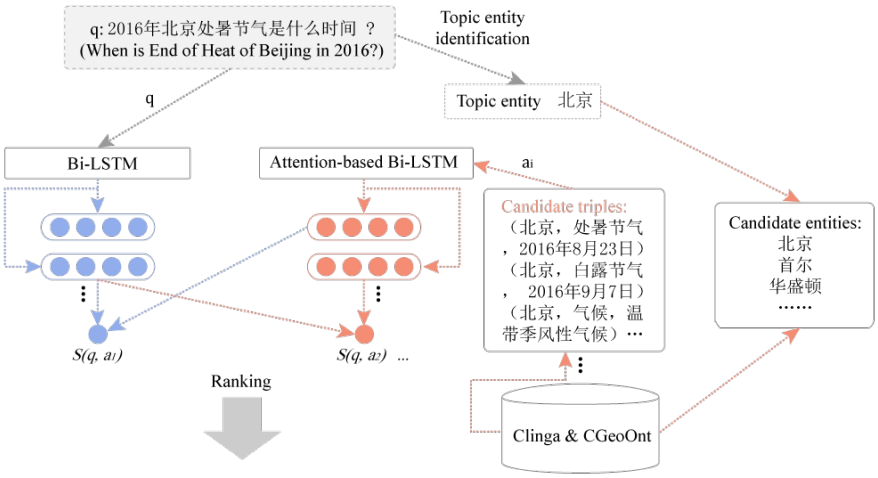


图 3.3: 基于注意力机制的地理知识库问答流程

（3）使用双向 LSTM 神经网络编码问题（问题词序列表示成词向量），使用基于注意力机制的双向 LSTM 编码候选答案三元组。

（4）计算问题与每个候选答案三元组的余弦相似度值，返回余弦值最大的三元组作为最终答案（有时返回多个最终答案，详见 3.4.1.4）。

3.4.1　地理知识库问答实现方法

此部分首先介绍候选答案实体三元组的生成方法，然后介绍本文表示问题、答案的基本模型 LSTM，再到其变种双向 LSTM 和结合注意力机制的双向 LSTM 模型，最后介绍模型的训练方法和问答最终答案的选择策略。

3.4.1.1 候选答案实体三元组生成

生成候选实体三元组需要识别问题中的主题实体，主题实体应当存在于本文所构建的知识库中，如果知识库中不存在此主题实体，则问题回答失败。

本文根据知识库构建实体概念词典以及实体概念别名词典，确定知识库能够回答的实体问题范围。候选答案实体三元组生成的关键是识别出问题中的主题实体，如下（1）-（3）为主题实体识别流程，（4）-（5）步为根据主题实体生成候选答案实体三元组：

（1）使用分词工具 jieba[7]将问题进行分词，生成问题的词序列。

（2）对词序列进行命名实体识别（Named Entity Recognition, NER）、词性标注（Part of Speech, POS）, 识别其中的命名实体、名词以及名词词组，本文称为候选主题实体集。

（3）遍历（2）步中生成的候选主题实体集，查询该候选实体是否在本文本体词典中，若存在，则该实体视作主题实体；若为本体词典中词的子串（不包含相等的情况），

[7]https://pypi.org/project/jieba/

亦将该实体视作主题实体；若所有候选主题实体都不存在于本体词典中，则此问题没有主题实体，此题目超出本文知识库的回答范围，回答失败。

（4）从本文地理知识库中获取跟主题实体一跳 1-hop 和二跳 2-hop 的实体集合作为候选答案实体集。

（5）最后，从知识库查询出候选答案实体集每个实体的三元组信息，构成目标候选答案实体三元组集合。

3.4.1.2 地理问题、候选答案表示

本节讲述如何将地理问题和候选答案文本序列表示成词向量（word embedding）形式。本文表示地理和候选答案的基础模型为循环神经网络（Recurrent Neural Network，RNN)，本节介绍使用 RNN 的变种 LSTM 模型来表示地理问题和答案，先介绍 RNN 表示序列数据的原理，后依次介绍使用基于基本的 LSTM 表示问题、答案，基于 Bi-LSTM 来表示问题答案和基于 Attention 的 Bi-LSTM 表示答案。

□ RNN 表示序列数据原理

RNN 的网络结构如图3.4所示，图左边部分表示 RNN 的循环结构图，RNN 的主体结构单元 A 处理来自当前时刻的输入 x_i 和上一时刻 A 输出的隐含状态。图的右边是左边图的展开形式。RNN 中每一层不仅输出 h_i 到下一层，而且同时还输出一个隐含状态（hidden state），该隐含状态表示当前层和之前所有层保存的信息，可以形象化地理解为当前到之前层的所有信息的综合记忆，下一层可以利用上一层输出的隐含状态信息。因此，RNN 此网络特征结构也使其擅长处理前后有依赖关系的序列数据。

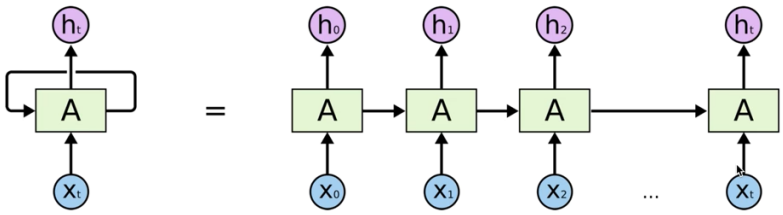


图 3.4: RNN 网络结构

□ 基于 LSTM 的问题、答案表示

在实际中，对于文本序列来说，循环神经网络较难捕捉两个时刻距离较大的文本元素（字或词）之间的依赖关系。LSTM 是 RNN 的变种之一，是一种常用的门控循环神经网络，可以解决 RNN 中的梯度消失、梯度爆炸问题。本文实现的 LSTM 为基于 Graves[49] 等人改进的 LSTM。给定输入问句，令其序列 **x**={x(1),x(2),⋯x(n)}，其中 x(t) 为问句中每个词的 d 维词向量表示，隐藏单元向量 h_t 遵循如下更新方式：

$$i_t = \sigma(\mathbf{W}_{ix}\mathbf{x}(t) + \mathbf{W}_{ih}\mathbf{h}(t-1) + \mathbf{b}_i) \tag{1}$$

$$f_t = \sigma(\mathbf{W}_{fx}\mathbf{x}(t) + \mathbf{W}_{fh}\mathbf{h}(t-1) + \mathbf{b}_f) \tag{2}$$

$$o_t = \sigma(\mathbf{W}_{ox}\mathbf{x}(t) + \mathbf{W}_{oh}\mathbf{h}(t-1) + \mathbf{b}_o) \tag{3}$$

$$\tilde{C}_t = \tanh(\mathbf{W}_{cx}\mathbf{x}(t) + \mathbf{W}_{ch}\mathbf{h}(t-1) + \mathbf{b}_c) \tag{4}$$

$$C_t = i_t \odot \tilde{C}_t + f_t \odot C_{t-1} \tag{5}$$

$$\mathbf{h}_t = o_t \odot \tanh(C_t) \tag{6}$$

此 LSTM 结构通过三个门，输入门 i、输出门 o、遗忘门 f 和记忆单元 c 来控制该时刻前的历史信息是否需要记忆、是否需要更新，从而更好的对历史信息进行保存，供下一时刻使用。公式中的 $\mathbf{W}_{ix}$、$\mathbf{W}_{ih}$、$\mathbf{W}_{fx}$、$\mathbf{W}_{fh}$、$\mathbf{W}_{ox}$、$\mathbf{W}_{oh}$、$\mathbf{W}_{cx}$、$\mathbf{W}_{ch}$、$\mathbf{W}_{cx}$、$\mathbf{W}_{ch}$ 是可学习的权重参数，$\mathbf{b}_i$、$\mathbf{b}_f$、$\mathbf{b}_o$、$\mathbf{b}_c$ 是可学习的偏移参数，$\mathbf{h}(t-1)$ 为上一时刻的隐含状态输出值，函数 σ 为 sigmoid 激活函数，tanh 为双曲正切函数作为激活函数，$\tilde{C}_t$ 表示长短期记忆中的候选记忆单元值，C_{t-1} 为上一时刻的记忆单元值，C_t 为当前时刻的记忆单元值，$\odot$ 为按元素乘法符。

本文首先通过分词将问题转化为词序列，此处记作 **q** = {x(1), x(2),⋯x(n)}, x(i) 表示问题词序列第 i 个词，然后查询词向量矩阵（初始为根据中文维基百科训练得到）获得每个词的初始词向量，最后将序列的词向量输入到上述 LSTM 模型，遵循其更新法进行训练，得到序列最终的词向量。

□ 基于 Bi-LSTM 的问题、答案表示

单向 LSTM 模型只考虑当前输入之前的信息，没有考虑其输入之后的信息，往往一个词的含义需要综合考虑此词的前后两部分词信息。因此，本文使用 Bi-LSTM，综合考虑当前输入的前向和后向信息，隐藏单元为前向 $\overrightarrow{h_t}$、后向单元 $\overleftarrow{h_t}$ 相连接，如公式（7 ）所示：

$$h_t = (\overrightarrow{h_t}, \overleftarrow{h_t}) \tag{7}$$

因此，本文基于 Bi-LSTM 的问答模型可以表示为图3.5问答模型所示。

该问答模型首先生成问题、答案词序列的初始词向量, 然后输入到 Bi-LSTM 网络进入训练，得到问题和答案的最终向量表示，最后词向量做 max pooling 后通过余弦相似性计算问题、答案向量矩阵的相似性。实验设置问题、答案的 Bi-LSTM 网络共享参数，Feng[50] 等人的研究表明两个网络共享参数较两个网络拥有各自不同参数性能更优。

□ 基于注意力机制的答案表示

与上述模型单独对问题和答案进行表示不同，此节使用一种基本的注意力机制模型，答案中每个词的向量生成均依赖问题，通过动态地将答案中更多的有效信息与问题相应关键信息对齐，可以更好的表示答案与问题之间的依赖关系。该注意力机制已在许多自然语言处理任务上取得不错效果，如机器翻译、事实型问答、句子摘要等。

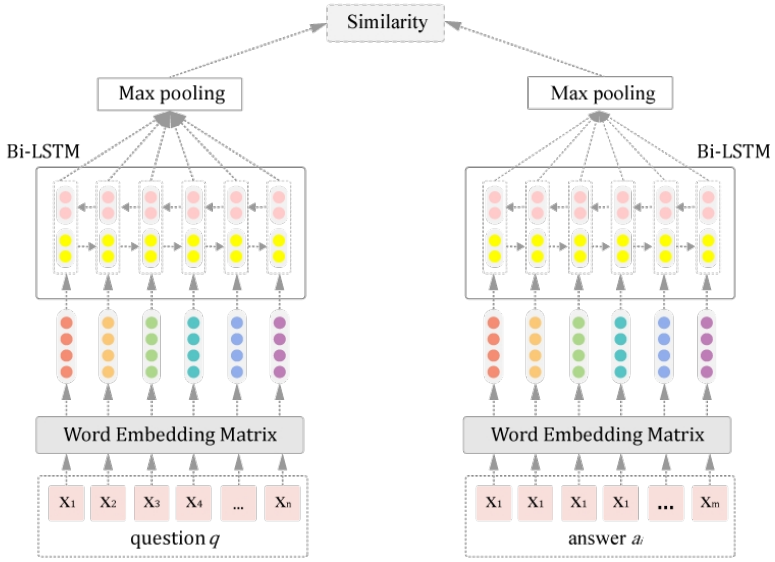


图 3.5: 基于 Bi-LSTM 的问题、答案表示模型

图 3.6 为本文基于注意力机制的答案表示模型，模型使用词级别的注意力，如图 3.6 中描述，在答案端 Bi-LSTM 网络输出进入 max pooling 层之前，每个输出会乘以一个基于问题的 Attention 值。时间 t 时刻，令答案端的 Bi-LSTM 输出向量为 $\mathbf{h}_a(t)$，问题词向量为 $\mathbf{o}_q$，答案中每个词向量 $\tilde{\mathbf{h}}_a(t)$ 更新如下：

$$\mathbf{m}_{a,q}(t) = \tanh(\mathbf{W}_{am}\mathbf{h}_a(t) + \mathbf{W}_{qm}\mathbf{o}_q) \tag{8}$$

$$s_{a,q}(t) = softmax(\mathbf{w}_{sm}\mathbf{m}_{a,q}(t)) \tag{9}$$

$$\tilde{\mathbf{h}}_a(t) = \mathbf{h}_a(t)s_{a,q}(t) \tag{10}$$

其中，$\mathbf{W}_{am}$、$\mathbf{W}_{qm}$ 为注意力机制参数矩阵，$\mathbf{w}_{sm}$ 为注意力机制参数向量。从直观上看，运用注意力机制表示答案相当于一定程度上将答案中与问题相对应关键词增加权重，给无关词减少权重，使更易于区分正确答案和错误答案。

时间 t 时刻，令答案端的 Bi-LSTM 输出向量为 $\mathbf{h}_a(t)$，问题词向量为 $\mathbf{o}_q$，答案中每个词向量 $\tilde{\mathbf{h}}_a(t)$ 更新如下：

其中，$\mathbf{W}_{am}$、$\mathbf{W}_{qm}$ 为注意力机制参数矩阵，$\mathbf{w}_{sm}$ 为注意力机制参数向量，$s_{a,q}(t)$ 为归一化后的注意力值。从直观上看，运用注意力机制表示答案相当于一定程度上增加答案中与问题相对应关键词权重，减少无关词权重，使正确答案和错误答案更易区分。

3.4.1.3 模型训练

本文采取问答常用的训练方式 pairwise training[51]，训练数据形式为（问题、正确答案、错误答案）。对于某个问题，本文选取一个正确答案作为正例，同时选取 k 个错误

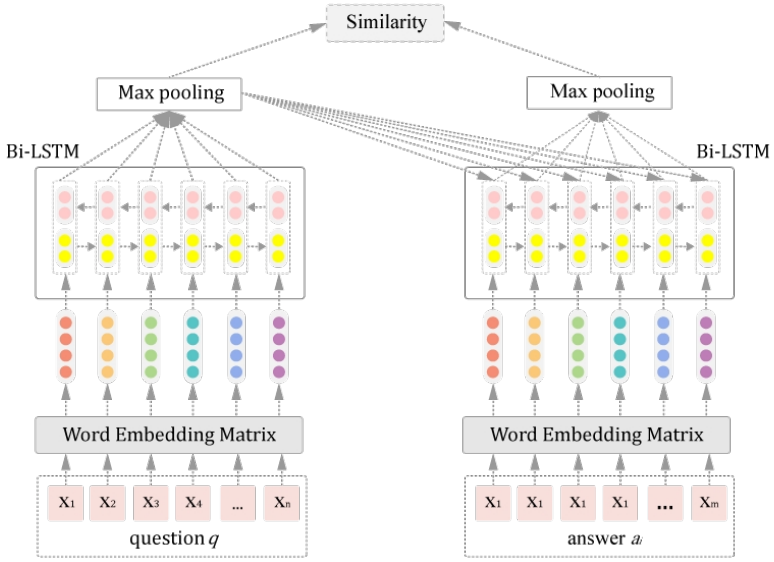


图 3.6: 基于 Attention 机制的问题、答案表示模型

答案作为负例，负例的选择一部分来自主题实体的非答案三元组，另一部分随机从其他问题的正确答案中选取。损失函数使用铰链损失函数（hinge loss），如下：

$$\mathcal{L}(q, a_+, a_-) = max\{0, m - S(q, a_+) + S(q, a_-)\} \tag{11}$$

m 为正确答案得分和错误得分差距，S(.) 为余弦函数，a_+ 为问题的正确答案，a_- 为问题的错误答案。目标函数为：

$$\min \sum_{q} \frac{1}{|P_q|} \sum_{a_+ \in P_q} \sum_{a_- \in N_q} \mathcal{L}(q, a_+, a_-) \tag{12}$$

其中 $|P_q|$ 为问题 q 的正确答案个数，N_q 为问题 q 的 k 个错误答案集合。本文采取随机梯度下降（stochastic gradient descent，SGD）学习算法。

3.4.1.4 最终答案选取策略

在测试阶段，问答模型给所有候选答案打分排序，选择出得分最高的答案三元组 S_{max}，考虑到一个问题的正确答案可能不止一种，因此，本文设置答案阈值，此阈值取损失函数中的 m，将得分与最高分相差不超过 m 的三元组也视作最终答案。如下：C_q 为问题 q 的所有候选答案集合，$\widehat{A}_q$ 为最终答案集合，$\hat{a}$ 为候选答案。

$$\mathcal{S}_{max} = \arg\max_{a \in C_q} S(q, a) \tag{13}$$

$$\widehat{A}_q = \{\hat{a} | S_{max} - S(q, \hat{a}) < m\} \tag{14}$$

3.4.2 地理问答实验

为了评估实验问答模型，本文使用 863 项目组构建的基础知识库，同时为了解决目前中文地理领域问题缺乏的问题，本文又构建了较大规模的地理问题测试集，详细在本章数据集小节 3.4.3 讲述。先介绍实验环境、参数设置，然后介绍实验结果以及实验分析，最后介绍地理数据集的生成过程。

3.4.2.1 实验环境

本文知识库问答实验是在微软云服务器（Microsoft Azure）上进行，具体硬件、软件环境如下所示：

硬件环境：

CPU: 4 × Intel(R) Xeon(R) CPU E5-2600 0 @ 2.20GHz

Memory: 14GB

软件环境：

系统：Ubuntu 16.04.4 LTS

编程语言：Python3.6.5 |Anconda

深度学习框架：Tensorflow 1.6.0

Python 编辑器：VIM -Vi IMproved 7.4 & Jupyter Notebook

知识库存储数据库：Virtuoso 7.2.4 Released

3.4.2.2 参数设置

实验词向量由 word2vec[52] 结合最新的中文维基百科语料训练得到，词向量维度为 300。实验采取随机梯度下降优化策略，实验尝试过不同的 m 值，如 0.1、0.2、0.3，最终选择 0.1 效果较佳。学习率初始值为 0.4，每 10 个 epochs 下降 50%，总共下降 4 次。剪枝参数 max_grad_norm 设为 5，dropout 取 1.0。实验 batch_size 取 50，问题负例个数 k 取 20，问题和答案的最大长度设置为 50，超出最大长度内容舍弃，双向 LSTM 隐藏单元长度设为 200。

3.4.2.3 评价指标

本文地理知识库问答实验采用知识库问答任务中常用的评价指标：平均倒数排序（Mean Reciprocal Rank，MRR ）、准确率 Accuracy@N[53]。如下为两个指标的定义：

$$MRR = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{rank_i} \tag{15}$$

公式（15）中 $|Q|$ 表示评估集中问题总数，$rank_i$ 表示当前问题的候选答案生成集中正确答案的位置，其中候选答案生成集是按照答案打分降序排序。若候选答案生成集中没

有包括正确答案，则 $rank_i$ 的值设为 0。

$$Accuracy@N = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\delta(C_i, A_i) \tag{16}$$

公式（16）中 $|Q|$ 表示评估集中问题总数，$\delta(C_i, A_i)$ 表示候选答案生成集中是否至少包含一个正确答案，若是，则 $\delta(C_i, A_i)$ 值为 1，否则，值为 0。

3.4.2.4 实验结果

实验使用 NLPCC2016[8]中 1000 个问题和本文从互联网抓取的 436 个地理问题（问题持续人工筛选中）进行测试，分别对本文方法章节中模型进行实验，问答评价指标使用 MRR 和 Accuracy@N，实验结果如表3.1所示：

表 3.1: 地理问答实验结果

Method	MRR	Accuracy@N
LSTM	0.809	0.847
Bi-LSTM	0.821	0.868
Bi-LSTM + Attention	**0.834**	**0.872**

3.4.2.5 实验分析

由表3.1中实验结果可知，Bi-LSTM 相对单向 LSTM 的 MRR、Accuracy@N 分别提高了 1.2%、2.1%，说明 Bi-LSTM 相比 LSTM 能表示问题能力更强。同时基于注意力机制的 Bi-LSTM 比单纯的 Bi-LSTM 两个指标分别提升 1.3%、0.4%，加入 Attention 后问答 MRR 指标提升比较明显，说明本文注意力机制可以更准确的区分比较相似的答案。

3.4.3　地理问答数据集构建

中文问答数据集 NLPCC2016 为开放百度百科类的问题，包含地理领域问题很少，并且问句组织形式单一、直接，因此无法有效地训练和测试本文地理问答模型，本文需要构建既包含专业地理知识、又问法多样的问题。

本文使用百度下拉框关键字推荐 API，将 10160 个核心地理知识三元组（s，p，o）分为（s，p），（s），（s，o），（p，o）四种搜索序列，分别获取百度下拉框推荐的问题列表，同时使用百度搜索 API 搜索这四种序列，获取每个序列搜索结果的前 10 个链接的标题，总共得到 246,767 个搜索结果。自动去掉一些重复、百度文库、百度百科词条标

[8]tcci.ccf.org.cn/conference/2016/

题等无效结果，最终得到 136680 个搜索结果，然后人工挑选其中可以当作地理问题的结果问题。

人工挑选地理问题时，还需要从本文地理核心知识库中找出可以回答该问题的实体三元组，无法从知识库中找出对应三元组的问题将被丢弃。目前已经人工挑选出 636 个地理问题，目前仍在继续人工标注中。本文将其中 200 个问题及其答案和 NLPCC2016 训练集中的 10000 个训练样例当作实验训练集。其余 436 个地理问题作为测试集的一部分。

根据地理核心三元组知识获取的多样化 web 地理问题举例如表3.2。

表3.2列举了地理知识三元组“（季风气候，生产优势，夏季高温多雨、雨热同期）”得到四种不同方式表达的问题（1）—（4），以及根据三元组得出问题的最终答案为“夏季高温多雨、雨热同期”。

（1）亚热带季风气候在发展农业生产方面有什么优势

（2）我国的季风气候对农业生产最有利的是?_ 作业帮

（3）季风气候的最大优点？

（4）温带季风气候在农业生产方面的显著优势是 _ 知道。

表 3.2: 地理问题集举例

知识库三元组	（季风气候，生产优势，夏季高温多雨、雨热同期）
web 问题 _1	亚热带季风气候在发展农业生产方面有什么优势
web 问题 _2	我国的季风气候对农业生产最有利的是?_ 作业帮
web 问题 _3	季风气候的最大优点？
web 问题 _4	温带季风气候在农业生产方面的显著优势是 _ 知道
问题答案	夏季高温多雨、雨热同期

3.5 本章小结

本章介绍了基于本体的地理知识问答的具体实现过程。主要包括对本文两个主要任务——地理本体知识库构建和基于地理知识库的问答的实现过程进行详细阐述。本章先介绍本文的任务以及本文的系统结构图；再介绍地理本体知识库的构建过程，包括如何构建地理本体 CGeoOnt，如何将基于百度百科构建的本体 Clinga 与 CGeoOnt 进行融合得到规模更大的地理本体知识库、如何存储和检索本文地理本体知识库；然后介绍基于地理本体知识库的问答模型及模型的训练方法，包括介绍基于 LSTM 表示地理问题和答案的问答模型、基于 Bi-LSTM 表示地理问题和答案的问答模型和基于 Attention 的 Bi-LSTM 表示地理答案的问答模型；最后介绍地理问答实验和地理问答数据集的构建过程，包括介绍地理问答实验的实验环境、实验参数设置、评价指标、实验结果及分析。

第四章　总结与展望

4.1　总结

本文为辅助解答高考地理题所尝试的工作，主要解决了在辅助解答地理高考题过程中的两个问题，第一个是缺乏高度结构化的地理核心知识库，第二个是地理问题表达多样而导致其理解困难。

为解决缺乏高度结构化的地理核心知识库的问题，本文以高中地理教科书和北京市近十年高考地理试题为知识源，通过自底向顶的本体构建思想，先从近十年高考题中提炼出地理核心考点，然后去地理教材中寻找可以解答核心考点的核心地理知识，以地理教科书中的知识组织框架为大体的本体知识层次组织，最后通过描述能力较强的本体语言 OWL DL 表示地理核心知识，得到高度结构化的地理本体 CGeoOnt。同时，本文还将 863 项目组以百度百科自动构建的中文地理本体 Clinga 与 CGeoOnt 融合得到目前中文地理领域规模较大、质量较高的地理本体知识库，用于辅助地理高考解题。

为解决地理问题表达多样而导致其理解困难问题，本文以 Clinga、CGeoOnt 为中文地理核心知识库，构建了一个对问题适应性较强的问答系统。该问答系统以 Attention-based Bi-LSTM 为问答模型，分别用两个共享参数的 Bi-LSTM 网络来表示问题和答案。同时，答案的表示结合答案序列对问题的注意力权重，模型使用有标记的问答对做训练，在从 web 收集到的多样性的地理问题数据集上，问答指标 MRR、Accuracy@N 分别达到 0.834、0.872 的较结果，当一个问题以不同方式提问时，本文模型同样具有较好的适应性。

本文构建的问答系统致力于辅助解答高考地理多选题，对于单实体单关系的地理问题，即使该问题提问方式形式多样，实验表明本系统仍可以比较好地解决，因此对于解答高考地理多选题有一定的实用价值。

4.2　未来展望

本文构建地理本体 CGeoOnt 采用的是人工构建的方式，从整体来看虽然构建的本体质量比较高，但是总体时间花费过大，六人标注团队一年半的时间仅仅标注两万多条地理知识三元组。因此，可以尝试使用自动或者半自动加人工的方式来构建地理本体，并且开发合适的地理知识自动或者半自动标注工具，提高地理标注效率。

再者，对于本文构建的地理知识库问答系统，从模型上看，对于答案的表示，除了本文结合问题中词级别的注意力机制，模型还可以结合地理题本身的特征，如地理问题答案类型、答案关系等，或者还可以结合地理本体知识库本身丰富的语义特征，将实体

的上下文类别语义信息添加到问题、答案的表示中，从这些方面更加完善地对问题、答案进行表示，从而提升问答的性能。

最后，本文从 web 收集的地理问题、答案数据集数量还很有限，并且问题的类别没有经过细致的分类划分。因此，可以尝试再扩大数据集的规模，将问题类型进行细分，使每种类型的问题均达到一定规模，使地理问答模型可以在不同类型问题上做相应的对比实验。

致谢

经过半年多的不懈努力，毕业论文的写作也将画上句号。一路走来遇到了很多艰难险阻，期间也得到老师、同学和朋友的指导与帮助。在此学位论文即将完成之际，我要向所有曾经给予我帮助的人表示最真诚的感谢。

首先，我要感谢我的导师——计算机科学与工程学院高志强教授。从论文的结构规划到论文的写作，高老师都给予我非常丰富的指导建议。除了学术论文上的指导，还感谢高老师身体力行的传授我许多科学的学习方法、高效的做计划习惯以及豁达的为人处事智慧，这些将使我受益终身。还要感谢我的校外导师王会方老师，王老师从工程的角度给我的论文提出了很多宝贵建议，促使我的论文不断完善。

其次，感谢我的所有同门朱曼、刘倩、鲁廷明、全志斌、归耀城、李雪莲、吴展鹏、吕永涛、王辰、倪朝曦、潘敬敏、司马强、王煜、王李荣、余云秀、刘金晶、范云龙、李斌、刘延栋、汪文涛、衣克买提、王延龙、刘颖、高翔、尹浩、尹俊、刘征等，尤其是鲁廷明师兄、全志斌师兄、李雪莲师姐、余云秀，感谢他（她）们在学术上给予的宝贵意见和帮助，使我的论文更加完善。

最后，感谢我的家人在此期间给予我的支持与信任，正是由于家人背后的默默付出，我才可以安心学习，并顺利完成学业。

毕业在即，今后的工作生活中，我定会谨记师长们的教诲，持续不断地努力、奋斗、不忘初心、追求卓越、不负众望!

张赏

2018 年 5 月于南京

参考文献

[1] Fujita, A., Kameda, A., Kawazoe, A., & Miyao, Y. Overview of Todai Robot Project and Evaluation Framework of its NLP-based Problem Solving[C].World History. 2014. 36.

[2] Cheng, G., Zhu, W., Wang, Z., Chen, J., & Qu, Y. Taking Up the Gaokao Challenge: An Information Retrieval Approach[C]. In IJCAI. 2016. 2479–2485.

[3] Hitzler, P., Krotzsch, M., & Rudolph, S. Foundations Of Semantic Web Technologies[M]. CRC press. 2012. 12–40.

[4] Zettlemoyer, L. S., & Collins, M. Learning context-dependent mappings from sentences to logical form[C]. In Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, 2009, 2: 976–984.

[5] Cai, Q., & Yates, A. Large-scale semantic parsing via schema matching and lexicon extension[C]. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, 2013, 1: 423–433.

[6] Berant, J., Chou, A., Frostig, R., & Liang, P. Semantic parsing on freebase from question-answer pairs[C]. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. 2013. 1533–1544.

[7] Dyer, C., Ballesteros, M., Ling, W., Matthews, A.,& Smith, N. A. Transition-based dependency parsing with stack long short-term memory. arXiv preprint arXiv:1505.08075.2015.

[8] Yao, X., & Van Durme, B. Information extraction over structured data: Question answering with freebase[C]. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, 2014, 1: 956–966.

[9] Bordes, A., Chopra, S., & Weston, J. Question answering with subgraph embeddings[C]. In Proceedings of EMNLP. 2014. 615–620.

[10] Dong, L., Wei, F., Zhou, M., & Xu, K. Question answering over freebase with multi-column convolutional neural networks[C]. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, 2015, 1: 260–269.

[11] Bordes, A., Usunier, N., Chopra, S., & Weston, J. Large-scale simple question answering with memory networks. arXiv preprint arXiv:1506.02075. 2015.

[12] Wong, Y. W., & Mooney, R. Learning synchronous grammars for semantic parsing with lambda calculus[C]. In Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics. 2007. 960–967.

[13] Yih, S. W. T., Chang, M. W., He, X., & Gao, J. Semantic parsing via staged query graph generation: Question answering with knowledge base[C]. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. 2015. 1321–1331.

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

[15] Zhang, Y., Liu, K., He, S., Ji, G., Liu, Z., Wu, H., & Zhao, J. Question answering over knowledge base with neural attention combining global knowledge information. arXiv preprint arXiv:1606.00979. 2016.

[16] Corcho, O., & Gómez-Pérez, A. A roadmap to ontology specification languages. In International Conference on Knowledge Engineering and Knowledge Management[C]. In Springer, Berlin, Heidelberg. 2000. 80–96.

[17] Bray, T., Paoli, J., Sperberg, C. Extensible Markup Language (XML) 1.0. W3C Recommendation. Feb, 1998. http://www.w3.org/TR/REC-xml.

[18] Lassila, O., Swick, R. Resource Description Framework (RDF) Model and Syntax Specification. W3C Recommendation. January, 1999. http://www.w3.org/TR/REC-xml.

[19] Brickley, D., Guha, R.V. Resource Description Framework (RDF) Schema Specification. W3C Proposed Recommendation. March, 1999. http://www.w3.org/TR/PR-rdf-schema.

[20] 高志强，潘越，马力. 语义 Web 原理及应用 [M]. 北京：机械工业出版社. 2009. 62–65.

[21] Bao, j., Calvanese D., Bernardo, C. G., et al. OWL 2 Web Ontology Language Document Overview (Second Edition). 2012. https://www.w3.org/TR/owl2-overview/.

[22] 岳丽欣, 刘文云. 国内外领域本体构建方法的比较研究 [J]. 情报理论与实践. In Natural Computation (ICNC), 11th International Conference on, 2016, 39(8), 119–125.

[23] Information Integration for Concurrent Engineering(IICE)[EB/OL]. 2015. http://www.kbsi.com/projects/iice?highlight=WyJpZGVmNSJd.

[24] USCHOLD M., GRUNINGER M. Ontologies: principles, methods and applications[J]. In International Conference on Computational Learning Theory. Knowledge Engineering Review, 1996, 11(2): 14–17.

[25] THAM K. D., FOX M. S., GRNINGER M. A cost ontology for enterprise modeling department of industrial engineering[C]. Proceedings of third Workshop on Enabling Technologies: Infrastructure for Collaborative Enterprises, Morgantown, WVWorkshop on Enabling Technologies. 2014. 111–117.

[26] FERNÁNDEZ M., GÓMEZ-PÉREZ, JURISTO N. METHONTOLOGY: from ontological art towards ontological engineering[J]. Sprng Symposium on Ontological Engineering of AAAI, 1997, (5): 33–40.

[27] Song, Y., Wang, H., & He, X. The KACTUS booklet version 1. 0. esprit project 814 [EB/OL]. 2015. http://www.swi.psy.uva.nl/prjects/NewKACTUS/Reports.html.

[28] 易利涛, 周肆清, 丁长松. 信息抽取中领域本体建模方法研究 [J]. 计算机技术与发展, 2011, (10): 23–27.

[29] 胡伟. 本体构建 [EB/OL]. 2014. http://www.docin.com./p-964406138.html.

[30] NOY, N. F., MC GUINNESS, D. L. Ontology development 101: a guide to creating your first ontology[J]. Knowledge Systems Laboratory, 2001, 32(1). 525–575.

[31] Lu, W., Cheng, J., Yang, Q. Question answering system based on web[C]. In Proceedings of the 2012 fifth international conference on intelligent computation technology and automation. IEEE Computer Society. 2012. 573–576.

[32] Bertola, F., & Patti, V. Ontology-based affective models to organize artworks in the social semantic web[J]. Information Processing & Management, 2016, 52(1): 139–162.

[33] Abacha, A. B., & Zweigenbaum, P. MEANS: A medical question-answering system combining NLP techniques and semantic Web technologies[J]. Information Processing & Management, 2015,51(5): 570–594.

[34] Pavlić, M., Han, Z. D., & Jakupović, A. Question answering with a conceptual framework for knowledge-based system development “Node of Knowledge”[J]. Expert systems with applications, 2015, 42(12): 5264–5286.

[35] Dalmas, T., & Webber, B. Answer comparison in automated question answering[J]. Journal of Applied Logic, 2007, 5(1): 104–120.

[36] Dragoni, M., da Costa Pereira, C., & Tettamanzi, A. G. A conceptual representation of documents and queries for information retrieval systems by using light ontologies[J]. Expert Systems with applications, 2012, 39(12): 10376–10388.

[37] Li, F., & Jagadish, H. V. NaLIR: an interactive natural language interface for querying relational databases[C]. In Proceedings of the 2014 ACM SIGMOD international conference on Management of data. ACM. 2014. 709–712.

[38] Zettlemoyer, L. S., & Collins, M. Learning to map sentences to logical form: Structured classification with probabilistic categorial grammars. arXiv preprint arXiv:1207.1420.2012.

[39] Kwiatkowski, J., ettlemoyer, L., Goldwater, S. & Steedman, M. Inducing probabilistic CCG grammars from logical form with higher-order unification[C]. In Empirical Methods in Natural Language Processing (EMNLP). 2010. 1223–1233.

[40] Berant, J., Chou, A., Frostig, R., & Liang, P. Semantic parsing on freebase from question-answer pairs[C]. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. 2013. 1533–1544.

[41] Bordes, A., Weston, J., & Usunier, N. Open question answering with weakly supervised embedding models[C]. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases. 2014. 165–180.

[42] Fader, A., Soderland, S., & Etzioni, O. Identifying relations for open information extraction[C]. In Proceedings of the conference on empirical methods in natural language processing. Association for Computational Linguistics. 2011. 1535–1545.

[43] Bahdanau, D., Cho, K., & Bengio, Y. Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473. 2014.

[44] Luong, M. T., Pham, H., & Manning, C. D. Effective approaches to attention-based neural machine translation. arXiv preprint arXiv:1508.04025. 2015.

[45] Rush, A. M., Chopra, S., & Weston, J. A neural attention model for abstractive sentence summarization. arXiv preprint arXiv:1509.00685. 2015.

[46] Zhang, Y., Liu, K., He, S., Ji, G., Liu, Z., Wu, H., & Zhao, J. Question answering over knowledge base with neural attention combining global knowledge information. arXiv preprint arXiv:1606.00979. 2016.

[47] Tan, M., Santos, C. D., Xiang, B., & Zhou, B. LSTM-based deep learning models for non-factoid answer selection. arXiv preprint arXiv:1511.04108. 2015.

[48] Hu, W., Li, H., Sun, Z., Qian, X., Xue, L., Cao, E., & Qu, Y. Clinga: bringing Chinese physical and human geography in Linked Open Data[C]. In International Semantic Web Conference.Springer. Cham. 2016. 104–112.

[49] Graves, A., Mohamed, A., and Hinton, G.. Speech recognition with deep recurrent neural networks[C]. In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2013. 6645–6649.

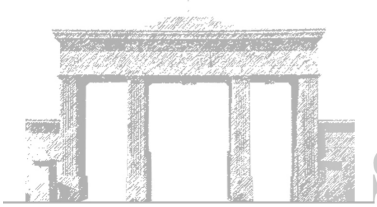
[50] Feng, M., Xiang, B., Glass, M. R., Wang, L., & Zhou, B. Applying deep learning to answer selection: A study and an open task. In Automatic Speech Recognition and Understanding (ASRU). Processing IEEE Workshop on. 2015. 813–820.

[51] Burges, C., Shaked, T., Renshaw, E., Lazier, A., Deeds, M., Hamilton, N., & Hullender, G. Learning to rank using gradient descent[C]. In Proceedings of the 22nd international conference on Machine. ACM. 2005. 89–96.

[52] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. Distributed representations of words and phrases and their compositionality[C]. In Advances in neural information processing systems. 2013. 3111–3119.

[53] Duan, N. Overview of the NLPCC-ICCPOL 2016 shared task: open domain chinese question answering[C]. In Natural Language Understanding and Intelligent Applications. Springer, Cham. 2016. 942–948.

心求至善



SOUTHEAST UNIVERSITY