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: 141466

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 2017 年 5 月 31 日
二级学科名称 计算机技术 学位授予日期 20 年 月 日
答辩委员会主席 陈国庆 评 阅 人 张志政 教授
匿名评阅人

2018 年 4 月 12 日

学校代码: 10286
分 类 号: TP311
密 级: 公开
U D C: 004.4
学 号: 141466

东南大学

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 2017 年 5 月 31 日

二级学科名称 计算机技术 学位授予日期 20 年 月 日

答辩委员会主席 陈国庆 评 阅 人 张志政 教授

匿名评阅人

2018 年 4 月 12 日

東南大學

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

本工作受到国家自然科学基金 61170165、61602260 和 61502095 的资助。

RESEARCH ON ENTITY LINKING APPROACHES BASED ON ACTIVE LEARNING

A Thesis submitted to

Southeast University

For the Academic Degree of Master of Engineering

BY

Wu Zhanpeng

Supervised by:

Prof. Gao Zhiqiang

School of computer science and engineering

Southeast University

2018/4/12

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性的任务。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。高考地理题的考点专业性很强，考点知识大都来自地理教科书章节知识，然而地理教科书是以文本文档的形式存在，无法准确表示出地理知识点之间的语义层次关系，因而也不适合当作计算机解答高考地理题的核心知识库。第二是地理问题表达形式多样，导致问题理解困难。地理问题中往往包含大量的无效信息，这些信息极易淹没问题核心考点信息，因而很难从大量干扰信息中精准地找出问题考点。针对以上问题，本文作了如下工作：

（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) 知识库 (Knowledge Base, KB)。该本体知识库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 CGeoOnt 与本体知识库 Clinga 进行本体融合，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，本文使用基于神经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，形成最终数据集。

关键词：　地理高考，本体，知识库问答，双向长短期记忆网络，注意力机制

Abstract

Entity linking is the task of determining the identity of entities mentioned in text. Supervised learning approaches and unsupervised learning approaches have been widely used in entity linking task in the past decade. However, only a few studies have been reported on accelerating model training and improving corpus construction. Active learning can contribute to interactively obtain optimum samples and provide them for annotators to conduct manual annotation according to the learning process. Meanwhile, it can reduce the quantity of the training samples as well as keep or improve model performance.

This thesis analyzes the characteristics of entity linking task, and use active learning approaches to handle model training and corpus construction task based on active learning.

The main contributions of this thesis include two aspects as follow:

(1) In consideration of supervised learning model of entity linking, this thesis reduces human annotating effort by using active learning, and proposes two approaches. One is an initial sample selection approach based on popularity, as known as sampling by popularity (SBP). The other is an iterative training sample selection approach based on comprehensive uncertainty and popularity, as known as sampling by uncertainty and popularity (SUP) . This way ensures representative of initial training sample in the initial sample selection stage and considers both uncertainty and representative of selected samples in the following stage of iterative sample training.

(2) To construct entity linking corpus, this thesis proposes an annotating approach based on active learning and unsupervised learning for improving annotation quality. In this way, the most informative samples of unlabeled mentions can be found for annotators to annotate while the precision rate of the whole corpus can be improved by propagating the evidence of labeled mentions.

Experiments in this thesis show two main points. One is that approaches of SBP and SUP can effectively accelerate the training process of entity linking model. The other is that approaches of annotation based on active learning and unsupervised learning can effectively improve the accuracy of annotating a silver-standard entity linking corpus on the premise of annotating fewer mentions.

Keywords: Entity Linking, Active Learning, Corpus Construction

目录

术语与符号约定

AL	Active Learning
EL	Entity Linking
EP	Evidence Propagation
NER	Named Entity Recognition
RW	Random Walk
SBP	Sampling by Popularity
SUP	Sampling by Uncertainty and Popularity

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通过高考。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 RobotProject）的人工智能项目，其最终目的是让此名为“Torobo”的“高考机器人”能够在 2021 年通过东京大学的入学考试 [1]。在 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地理高考多选题的一些尝试，图 1 所示为 2016 年上海地理高考多选题:

（**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

题中划线部分为此题问题，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，可得此题答案选“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）”，并且我们还需要知道“东南亚”的类别是一些国家的集合，“泰国”的类别属于东南亚国家。

鉴于以上分析，解答高考地理题多选题一般包含两步，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，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知识三元组。

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，描述概念的类别信息，描述各概念之间的相互作用关系。地理核心知识通过三元组（主、谓、宾形式元祖）形式得以更精炼的表示，地理核心概念层次关系明显，更适合作进一步的推理。

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，系统需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 [4-7] 和基于信息检索 [8-11]。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 上的问答实验表明，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，基于神经网络的模型只需将问题和答案分别表示成语义向量，然后计算向量相似性即可获得最相似的候选答案。问句和答案的向量表示是基于神经网络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，一个地理三元组往往可以成为多个问题的答案。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。虽然两个问题在问题表述不一样，但其问题核心均考察“季风气候的生成优势”，因此相对答案三元组而言它们是等效的。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准确的，至少是不能表示问题和答案之间关系的，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可以与问题中重要信息对齐，这样也减弱了问题中无效信息的影响，可以获得更好的问答效果。

1.2 研究内容

本文为辅助解答高考地理多选题所做的工作，本文核心内容为从构建的地理核心知识库中找出可以回答所求地理问题的知识三元组。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、高可用性的地理知识库。

同时，本文研究如何更准确的根据表达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，便于解题组根据此三元组作进一步的答案推理，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，构建了中文地理本体 CGeoOnt 知识库。本体知识库的构建使用 OWL 本体语言，将地理教材中的核心考点概念属性、核心概念之间的关系形式化表示。并且，运用启发式规则将 CGeoOnt 与本体 Clinga 进行融合，采取人工做最终的融合校验，形成更综合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知识库问答模型。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，使模型辨别答案能力更强。

（3）为解决中文地理问答模型在训练和测试中数据集缺失问题，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，经机器半自动筛选和人工筛选出有效的问题。最后，人工从本文构建的地理知识库中选择能回答这些问题的三元组知识，形成最终的问题、答案对数据集。

1.3 论文组织

本文共分为五章，各章的主要内容如下：

（一）第一章主要介绍相关的基本概念、应用背景，以及研究内容和论文的组织结构。

（二）第二章主要介绍实体链接任务和主动学习方法的研究现状。介绍了实体链接任务的基本概念以及常用的处理方法，分析了目前实体链接任务方法各自的优缺点；还介绍了主动学习方法的概念、常用的方法、以及它在其它机器学习任务中的应用。

（三）第三章首先介绍基于主动学习的实体链接模型的训练方法。然后介绍经过改进的主动学习方法，提出了基于流行度的初始样本选择方法以及综合不确定度和流行度的迭代训练样本选择方法。并通过实验验证了本文所提方法的有效性。

（四）第四章首先介绍基于主动学习的实体链接银标准语料构建方法。然后提出待标注样本的选择方法，以及标注结果的证据传播方法。并通过实验验证银标准语料库构建效率的提升效果。

（五）第五章对全文进行总结，指出本文的创新点和不足之处，并对未来的研究进行展望。

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；其次介绍问答的基本方法，包括基于语义解析的方法和基于信息检索的方法；最后介绍基于本体的地理知识问答方法。

2.1　本体研究

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]。运用本体可以很好的表示出领域中核心知识概念的语义信息和知识概念之间的相互关系。通过其 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的表示出来，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。本体的大多少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，使不同个体可以很好的进行区分或者是可以建立某种关联。

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，或者是有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、李四也为人的个体。

（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。

本体中类的成员（包括个体、类别）没有限制为互斥关系，因为通常情况下，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，一个类可以包含其他类或者被其他类包含，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出其自身的特征，还可以表示出其与其它知识的关联，而且这种关联是非常接近人类的概念思维，所以知识建模非常直观。

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【上标】。同样，属性与类结构类似，具有子属性层次。

属性另外一个重要特点是，属性含有定义域 (domain)、值域 (range) 限制 (restriction)。运用属性的这一限制我们可以对属性两边的个体做相关的类别推理。如根据申明的 *hasWife* 关系，可以推导出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连接在一起的。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也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、属于关系和子类型关系。这些关系都是表达的一种上下位的关系，特别是其中的 *is-a-subclass-of* 关系，它体现出一中很强的分类学思想，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”，显然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。这种特定领域的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 本体表示

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【14】。其中，由 W3C 推荐的 XML、RDF、RDFS 以及 OWL 使用最为广泛。

XML(Extensible Markup Language)【14 中的引用】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基本要素。RDF(Resource Description Framework)【】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) 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【高老师书】。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)，OWL 定义了逻辑类的关系表示，即提供了针对逻辑与、或、非的关系表示，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的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, 为了满足不同用户对本体的建模需求，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，其推理复杂度也逐渐提高。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，但它的推理最高效，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，比 OWL Full 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

时其支持有效的推理，因此在既对本体语言表达能力要求高，又需要保证推理的可判定性情景时，可以选择此本体语言；OWL Full 在这三种语言中表达能力最强，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，应该选择 OWL Full 语言。

2.3 实体链接任务

2.3.1 任务描述

实体链接任务，即在给定包含实体集 E 的知识库和包含指称项集合 M 的文本的前提下，将文本中的每一个指称项 $m \in M$ 链接到知识库中的对应实体 $e \in E$。这里所提到的指称项 m 是指文本中的一段字符串序列，称之为命名实体，通常是人名、地名、机构名。一些研究者也将实体链接任务称为命名实体消歧（Named Entity Disambiguation, NED）。按照所处理的语言种类区分，实体链接任务还可分为单语种的实体链接以及跨语种的实体链接[13]，本文主要研究英文实体链接任务。

通常，实体链接任务首先需要经过命名实体识别阶段，在该阶段，文本中的人名、地名、机构名等命名实体会被命名实体识别器识别出来，并划分出字符边界。命名实体识别经过几十年的研究，已经有不少成熟的研究成果[14;15;16]。

在工程实现中，也可借助各种开源的命名实体识别工具包，例如 Stanford NER[1]、OpenNLP[2]、LingPipe[3]等。由于命名实体识别并非本文重点，这里不再赘述。

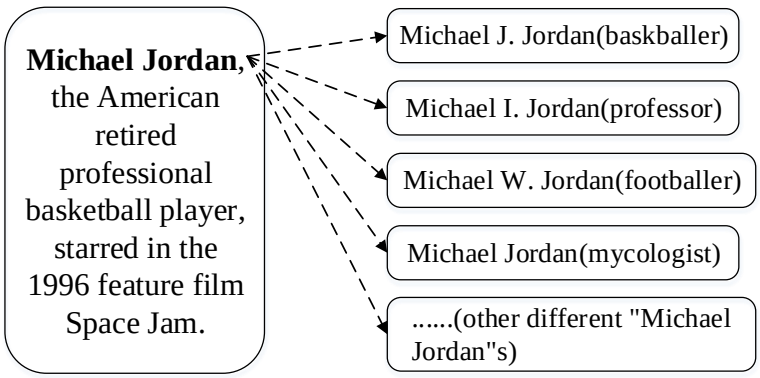


图 2.1: 实体链接任务的一个例子，文本中粗体标志的就是指称项。

图2.1展示了实体链接任务的一个例子。图中左侧是包含指称项“Michael Jordan”的文本，图中右侧是指称项“Michael Jordan”在知识库中可能指向的具体实体，例如 NBA 球星 *Michael J. Jordan*、伯克利机器学习教授 *Michael I. Jordan*、足球运动员 *Michael W.*

[1] http://nlp.stanford.edu/ner/

[2] http://opennlp.apache.org/

[3] http://alias-i.com/lingpipe/

Jordan、菌类学家 *Michael Jordan* 等。实体链接系统需要借助指称项上下文与候选实体在知识库中文本的相似度等特征，将指称项链接到其对应的目标实体。在该例子中，目标实体就是 NBA 球星 *Michael J. Jordan*。

一般来说，实体链接系统分为以下两个模块：

- 候选实体生成模块

  在这个模块中，实体链接系统会为文本中包含的每个指称项 $m \in M$ 生成出它在知识库中可能指向的候选实体集 E_m。经过多年的研究，研究者已经提出了大量候选实体生成方法。

- 候选实体排序模块

  该模块是实体链接系统中最重要的模块。在大多数情况下，指称项 m 都会对应多个候选实体 $e \in E_m$。实体链接系统需要从指称项 m 的候选实体集 E_m 中选出可能性最大的目标实体 e^*。处理候选实体排序的方法很多，可以划分为有监督学习方法和无监督学习方法。

2.3.2　候选实体生成

候选实体生成的方法很多，Hachey 等人[17] 认为能否生成包含目标实体的候选实体集对实体链接的成功与否至关重要。目前在实体链接领域中，使用最广泛的候选实体生成方法是基于词典的候选实体生成方法。

基于词典的候选实体生成方法被多种实体链接系统[18;19] 所采纳。以 Wikipedia 为例，该方法抽取了知识库中实体页（Entity pages）、重定向页（Redirect pages）、消歧页（Disambiguation pages）、文本超链接（Hyperlinks in articles）等结构化信息，通过这些信息构建了离线形式的词典 D，该词典 D 包含了指称项和指称项可能指向的候选实体。这里的指称项是知识库中实体的多种表述，可以是实体的标题名、缩写名、别名、拼写名等。

实质上，词典 D 即键值对 $\langle key, value \rangle$ 的集合，其中 key 列是指称项 k，$value$ 列是指称项 k 对应的候选实体集 $k.value$。本文可以借助 Wikipedia 知识库中的以下特征抽取得到词典 D：

- 实体页（Entity page）

  知识库中每一个实体都包含一个描述该实体的实体页，通常，这个实体页的标题就是该实体被最常用来表述的名字，例如实体页标题“Michael Jordan”就可以用来表示前 NBA 公牛队球员 *Michael Jordan* 这个实体。在这个例子里，标题 $k =$ “$Michael\ Jordan$” 被添加到词典 D 的 key 这一列，对应的实体 $k.value =$ “$Michael\ Jordan$” 被添加到词典 D 的 $value$ 这一列。

- 重定向页（Redirect page）

  在很多情况下，不同的表述可能指向同一个实体，对于这类情况，知识库会采用重定向的方式来处理。例如，对于标题为“Michael Jeffrey Jordan”的实体页，会重定向到标题为“Michael Jordan”的实体页。在此例中，标题 $k =$ “$Michael\ Jeffrey\ Jordan$” 被添加到词典 D 的 key 这一列，对应的实体 $k.value =$ “$Michael\ Jordan$” 被添加到词典 D 的 $value$ 这一列。

- 消歧页（Disambiguation page）

  在知识库中，不同的实体可以使用相同的表述。消歧页用来展示某个实体表述可以指向的实体集。例如，“Michael Jordan”这个实体表述，在消歧页“Michael Jordan（disambiguation）”中，包含 13 个指向不同人物的实体，有 NBA 球员、足球运动员、机器学习教授、爱尔兰政治家等。在这个例子里，标题 $k =$ “$Michael\ Jeffrey\ Jordan$” 被添加到词典 D 的 key 这一列，对应的 13 个实体 $k.value$ 被添加到词典 D 的 $value$ 这一列。

- 文本超链接（Hyperlinks in articles）

  知识库的文本中一般都会包含很多指向其它实体页的超链接，这些超链接的锚文本通常是所链接到的实体的标题或者别名。这类锚文本可以作为目标实体的一种指称项表述。例如，在“Chicago Bulls”这个实体对应词条的文本中，包含这样一段文本，“For his efforts, **Jordan** was named NBA Most Valuable Player.”，加粗的文本“Jordan”有指向实体页“Michael Jordan”的超链接。在这个例子里，标题 $k =$ “$Jordan$” 被添加到词典 D 的 key 这一列，对应的实体 $k.value =$ “$Michael\ Jordan$” 被添加到词典 D 的 $value$ 这一列。

表 2.1: 候选实体词典的例子

k(Name)	$k.value$(Entity)
Michael Jordan	*Michael Jordan* *Michael Jordan (footballer)* *Mike Jordan (racing driver)* *Michael I. Jordan* *Michael Jordan (Irish politician)* ... ...

通过以上的几种方式，可以构建出候选实体词典 D。在后续阶段，只要给出指称项 m，实体链接系统就能通过查找词典的方式得到 m 对应的候选实体集 E_m。表2.1是候选

实体词典中指称项为“Michael Jordan”的一个例子，根据 k 列的指称项，可以在词典 $k.value$ 列中查到“Michael Jordan”对应的所有可能指向的候选实体。一般来说，为了提高实体链接系统的性能，需要尽可能保证目标实体被包含在候选实体集中。

2.3.3　候选实体排序

对于给定的指称项 m，通过基于词典的候选实体生成方法获得候选实体集 E_m 后，候选实体集 E_m 一般包含多个候选实体，即 $|E_m| > 1$，例如“Michael Jordan”这个指称项对应的候选实体集就包含 *Michael Jordan*、*Michael Jordan (footballer)*、*Mike Jordan (racing driver)* 等多个候选实体。Shen 等人[20] 通过调研发现，TAC-KBP2010 数据集中平均每个指称项指向 12.9 个候选实体，TAC-KBP2011 数据集中平均每个指称项指向 13.1 个候选实体。实体链接系统需要对 E_m 中的候选实体做排序，将排序最靠前的实体作为最佳预测实体。候选实体排序模块是实体链接系统中的重要模块。研究者对候选实体排序方法在监督学习方法和无监督学习方法中均有做相关工作。

2.3.3.1　二分类法

二分类法（Binary Classification Methods）是处理实体链接任务中候选实体排序问题的一种简单高效的有监督学习方法。在给定一个指称项 m 和其对应候选实体集中的某一个候选实体 $e_i \in E_m$ 的样本对 $\langle m, e_i \rangle$ 后，候选实体排序模块的工作是判断 e_i 是否为 m 的目标实体。如果 e_i 是 m 的目标实体，则将样本对 $\langle m, e_i \rangle$ 划分为正例，否则将其划分为反例。

对于给定的样本对 $\langle m, e_i \rangle$，需要通过特征提取算法提取特征向量，然后将该向量作为模型的输入，模型的输出为样本的分类类型，即正例和反例。在基于二分类法的候选实体排序模型的训练阶段，需要提供足够的带标注的样本对 $\langle m, e_i \rangle$ 作为训练集，如果 e_i 就是目标实体，则标注为正例，否则标注为反例。当完成模型训练以后，给定未标注的样本对 $\langle m, e_i \rangle$，提取特征输入训练好的模型，模型返回分类结果，根据输出结果是正例还是反例判断 e_i 是否是 m 的目标实体。

二分类法的模型有很多种选择。Xiaohua 等人[21]、Jinlan 等人[22]、Yahui 等人[23] 使用支持向量机模型（Support Vector Machine, SVM）来处理二分类的实体链接问题。支持向量机是一种特征空间上的间隔最大化的分类器。另外，支持向量机的核函数（Kernel Function）能够将非线性可分的输入映射到高维线性可分的特征空间。大量研究表明，支持向量机特别适合处理二分类问题。其它可用于处理二分类问题的模型还包括朴素贝叶斯模型（Naive Bayes）、逻辑斯谛回归模型（Logistic Regression）、K-近邻模型（K-Nearest Neighbors）等。

2.3.3.2 排序学习法

排序学习法（Learning to Rank Methods）也是一种用于处理实体链接任务的监督学习方法，该方法最早被用来解决信息检索领域中的网页排序（Page Rank）问题。相比于二分类法，排序法克服了二分类法训练样本集正例、反例数量不平衡的问题，另外，当某个指称项的多个候选实体样本对都被判断为正例时，需要通过其它方法从这些候选实体中选出最有可能是目标实体的实体。与二分类法不同，排序学习法会对给定指称项对应的所有候选实体进行评分，然后根据评分对候选实体进行排序，最后选择评分最高的候选实体作为预测的目标实体。排序学习法主要可以分为三种方法，分别是基于数据点的方法（Pointwise）、基于数据对的方法（Pairwise）和基于列表的方法（Listwise）。

- 基于数据点的方法

  David 等人[24] 将基于数据点的方法用于网页排序任务。在实体链接任务中，基于数据点的方法会以候选实体为粒度，对指称项 m 和其中一项候选实体 $e_i \in E_m$ 的链接置信度进行计算。模型的输入是单个候选实体样本对 $\langle m, e_i \rangle$，模型的输出可以是回归值，也可以是分类值。例如，对于 $\langle m, e_i \rangle$，如果输出的是回归值评分，则选出 $e_i \in E_m$ 中评分最高的候选实体作为预测目标实体。

表 2.2: Pointwise 模型的例子

	基于数据点的方法	
	回归	分类
输入空间	$\langle m, e_i \rangle$	
输出空间	实数	分类
Hypothesis Space	评分函数 $f(\langle m, e_i \rangle)$	
损失函数	回归损失	分类损失
	$L(f; \langle m, e_i \rangle, y_j)$	

  如表2.2所示，对于回归值来说，无论是线性回归还是逻辑斯蒂回归，最后的输出都是一个实数，并可以对每个候选实体对应的实数进行排序。对于分类值来说，可以得出无序的实体类别及其置信度，即该实体是不是目标实体，以及是目标实体的置信度。

- 基于数据对的方法

  与基于数据点的方法不同，基于数据对的方法以指称项的候选实体对为粒度，对指称项 m 及其候选实体对 $e_i \in E_m$ 和 $e_j \in E_m$ 的链接置信度进行计算。模型的输入是由指称项和两个候选实体构成的三元组 $\langle m, e_i, e_j \rangle$。模型的输出是一个二分类标

签。例如，当候选实体 e_i 比候选实体 e_j 更接近目标实体时，模型输出为 1，反之则模型输出为-1。基于数据对的方法包括 RankBoost[25]、RankSVM[26]、RankNet[27]。

- 基于列表的方法

  基于列表的方法不再单独考虑某一个候选实体或者某一对候选实体，而是同时考虑一组候选实体，主要通过直接优化候选实体的评价方法和定义损失函数两种方法实现。基于列表的方法的主要模型包括：AdaRank[28]、SVM-MAP[29]、ListNet[30]、LambdaMART[31] 等。

  Cao 等人[30] 提出 ListNet 来描述 Listwise 的损失函数，该损失函数的定义是模型计算所得的候选实体排序和真实候选实体排序之间的差异程度，训练过程中通过算法将该差异最小化。训练完毕得到模型后，将候选实体排序问题转化为概率分布问题，用“交叉熵”来衡量计算得到的候选实体与真实排序的差异，通过最小化该差异来完成排序任务。

2.3.3.3　基于图的协同推断

Han 等人[18] 提出的基于图的协同推断方法（Collective Graph-Based Methods）是一种用于处理实体链接任务的无监督学习方法。该方法不仅考虑到了单个指称项和候选实体的上下文相似度，并将其作为局部相似度，还基于同一文档内不同指称项指向的实体具有关联性这一特点，借助知识库中实体的链接关系，综合考虑了实体之间的语义关联度，并将其作为全局相似度。在这两种相似度的基础上，构造用有向图表示的推理图，并基于图的协同推断方法处理候选实体排序问题。

构造出推理图后，可以通过协同推断算法，利用图模型中的随机行走（Random Walk）算法[32] 来确定同一文档中的各个指称项对应的最佳候选实体。Han 等人[18] 的实验表明，综合考虑了全局相似度这一特征后，在 IITB 数据集[4]上实验得到的 F1 值是 73%，实体链接系统性能得到了显著提升。

2.3.4　评测资源

评测资源主要由知识库和文本语料构成。本节将介绍目前可用的实体链接任务相关评测资源。

（1）目前在实体链接任务中常用的知识库主要如下：

- 维基百科[5]。该知识库是由非营利组织维基媒体基金会负责营运。2015 年 11 月英文版维基百科包含 400 万个实体，包括 832000 个人物、639,000 个地点、209,000 个机构等。

[4]http://www.cse.iitb.ac.in/ soumen/doc/

[5]http://www.wikipedia.org

- Knowledge Graph[6]。该知识库当前最新版本大约包含 5.7 亿个实体。
- TAC 会议的 KBP 任务发布的基于维基百科的知识库，大约包含 818741 个实体。

（2）目前在实体链接任务中常用的评测语料库主要如下：

- AIDA[33]。该数据集的语料来自英文新闻文本，并在 CoNLL'03[34] 的基础上做了实体链接标注。数据集包含 1393 篇文章，34956 个指称项目，但是 6141 个指称项对应的实体无法与维基百科词条对应，因此可用指称项个数为 28815 个。
- TAC-KBP2010[35] 发布的实体链接语料库。该语料主要来源于英文新闻以及 Web 文本，主要包含人物指称项 1877 个，机构指称项 3960 个，地理政治指称项 1817 个。

2.4 主动学习

监督学习模型被广泛用于分类问题，但是所有基于监督学习的分类模型都需要使用带标注的样本集对模型进行训练。对未标注的样本进行人工标注费时费力，并且训练样本可能存在重复，因此没有必要对这类样本做重复标注。主动学习能够根据学习进程，选择最佳学习样本交由人工标注。主动学习的样本选择过程主要分为两个阶段，分别是初始训练样本选择阶段和迭代训练样本选择阶段。

第一阶段，主动学习器构造一个规模较小的初始带标注的样本集，用于训练一个初始模型。第二阶段，在模型迭代训练过程中，主动学习器能主动选择包含信息量大的未标注样本交由专家标注，然后将这些带标注的样本加入到训练集，从而在保证模型泛化能力的同时，减小人工标注的工作量。

2.4.1 基于不确定度的样本选择策略

基于不确定度的主动学习算法采用了不确定度采样（Uncertainty Sampling）的方式选择待标注样本。该采样方式由 Lewis 等人[36] 于 1994 年提出，基于最大化人工标注效率的原则，尽量选择因置信度较小而更可能被错误分类的样本点，交由人工标注，以此加快模型的训练速度。

以基于 SVM 的二分类问题为例，样本点 x_i 到分类超平面的距离可由公式2.1计算：

$$f(x_i) = \sum_{j=1}^{n} \alpha_j y_j K(x_j, x_i) + b \tag{2.1}$$

公式2.1中 $K(x_j, x_i)$ 是 SVM 的核函数，代表样本点 x_i 和样本点 x_j 的相似度。根据样本点到分类超平面的距离，可以预测样本分类的置信度。离分类超平面越近，分类置

[6]https://www.google.com/intl/es419/insidesearch/features/search/knowledge.html

信度越低，越容易被主动学习器选中。反之，则分类置信度越高，越不容易被主动学习器选中。

基于池的主动学习方法[37]在目前研究中使用最为广泛。该方法的关键步骤是，在待标注样本选择的过程中设计算法对未标注样本的信息量进行定量分析。对于二分类问题来说，距离分类超平面最近的样本点就是分类置信度最接近 0 的样本点，这些样本点的分类不确定度相对较高。但是对于分类标签个数大于 2 的多分类问题来说，仅根据分类置信度是否接近 0 已无法区分样本的不确定度。对于多分类问题来说，不确定度的度量方式主要有三种。

（1）置信度。基于置信度的不确定度度量方式是用于度量不确定度最基本的方式。对于某一个样本点，其对应的所有可能的分类标签都会被赋予相应的置信度，将其中置信度最大的标签作为预测最佳标签，并将此标签的置信度作为该样本被正确分类的置信度。

$$\begin{aligned} x_{LC}^{*} &= \underset{x}{\operatorname{argmin}}\, P_{\theta}(\hat{y}|x) \\ &= \underset{x}{\operatorname{argmax}}\, 1 - P_{\theta}(\hat{y}|x) \end{aligned} \tag{2.2}$$

在公式2.2中，$y^{*} = \operatorname{argmax}_{y} P_{\theta}(y|x)$，即在当前分类器 θ 下，样本 x 的最佳预测分类标签。基于该度量方式，最佳预测分类标签的置信度越低，则分类的不确定度越高，这些样本点更应该被选择出来进行人工标注。该度量方式可由 0-1 损失（0-1 loss）来解释。该度量方式的缺点是只考虑了最佳分类标签的置信度，而丢弃了其他分类标签的置信度的分布情况。

（2）间隔。基于间隔的不确定度度量方式以最佳预测分类标签置信度和次佳预测分类标签置信度的差值作为考量因素。

$$\begin{aligned} x_{M}^{*} &= \underset{x}{\operatorname{argmin}}[P_{\theta}(y^{*1}|x) - P_{\theta}(y^{*2}|x)] \\ &= \underset{x}{\operatorname{argmax}}[P_{\theta}(y^{*2}|x) - P_{\theta}(y^{*1}|x)] \end{aligned} \tag{2.3}$$

在公式2.3中，y^{*1} 和 y^{*2} 分别是在当前分类器 θ 下，样本 x 的最佳预测分类标签和次佳预测分类标签。该度量方式通过同时考虑两个分类标签的置信度，在一定程度上克服了前一种度量方式的缺点。基于该度量方式，间隔越大的样本点，分类器越容易从最佳的两个分类标签中区分出最佳预测样本点，这种样本的分类置信度就较高。因此，主动学习器应该选择间隔较小的样本点，这些样本点的分类置信度较低，对模型的训练更有帮助。

（3）熵值。在主动学习方法中，目前使用较广泛的不确定度度量方式是熵值度量法，熵是用来描述模型对样本点分类时，分类结果混乱程度的一项指标。熵值的计算方法如公式2.4所示。

$$\begin{aligned} x_M^* &= \underset{x}{\text{argmax}}\, H_\theta(Y|x) \\ &= \underset{x}{\text{argmax}} - \sum_y P_\theta(y|x) \log P_\theta(y|x) \end{aligned} \tag{2.4}$$

通过公式2.4，可以看出，基于熵值的不确定度度量方式考虑了样本点 x 对应的所有可能的分类标签的分类置信度，从而使该方法可以描述样本分类结果的整体分布情况。对基于熵值的不确定度度量方式可以由对数损失（Logarithmic Loss）来解释。熵值越大的样本点，分类标签置信度的分布越混乱，样本点的分类不确定度越大。反之，分类标签置信度的分布越集中。极端情况，某一分类标签的置信度为 1，其他分类标签的置信度均为 0，则熵值为 0，这时候分类不确定度是最低的。因此，主动学习器应该选择熵值较大的样本点对模型进行重训练。

2.4.2 基于委员会的主动学习算法

与基于单个模型的主动学习算法不同，Seung 等人[38] 提出的基于委员会的主动学习算法会选择一定数量的分类模型，组成分类委员会。在待标注样本选择过程中，委员会中的各个模型分别对样本点进行分类。各委员会模型分类预测结果最不一致的样本点就是分类不确定度最高的样本点，主动学习器需要选择这些样本点进行人工标注并将其加入训练集对委员会中的所有模型进行重训练。

与基于单个模型的主动学习算法相比，基于委员会的主动学习算法的不同之处在于它包含多个分类器模型，并且在选择待标注样本时，选择的依据是委员会对某一个样本点分类的分歧度，分歧度越大的样本点，分类置信度越低，越应该被主动学习器选择。委员会分歧度的度量方式主要有以下两种。

（1）投票熵（Vote Entropy）。投票熵的定义如下：

$$x_{VE}^* = \underset{x}{\text{argmax}} - \sum_y \frac{vote_{\mathcal{C}}(y,x)}{|\mathcal{C}|} \log \frac{vote_{\mathcal{C}}(y,x)}{|\mathcal{C}|} \tag{2.5}$$

公式2.5中，y 表示所有可能的样本分类标签。$vote_{\mathcal{C}}(y,x) = \sum_{\theta \in \mathcal{C}} 1_{\{h_\theta(x)=y\}}$，即预测分类标签为 y 的分类器个数，$|\mathcal{C}|$ 是委员会中分类器的个数。在此基础上，研究者们提出了软投票熵的定义：

$$x_{SVE}^* = \underset{x}{\text{argmax}} - \sum_y P_{\mathcal{C}}(y|x) \log P_{\mathcal{C}}(y|x) \tag{2.6}$$

公式2.6中，$P_{\mathcal{C}}(y|x) = \frac{1}{|\mathcal{C}|} \sum_{\theta \in \mathcal{C}} P_\theta(y|x)$。软投票熵考虑了委员会中每个分类器的分类置信度。

（2）KL 散度（Kullback-Leibler Divergence）。KL 散度的定义如下：

$$x^*_{KL} = \underset{x}{\text{argmax}} \frac{1}{|\mathcal{C}|} \sum_{\theta \in \mathcal{C}} KL(P_\theta(Y|x)||P_{\mathcal{C}}(Y|x)) \tag{2.7}$$

$$KL(P_\theta(Y|x)||P_{\mathcal{C}}(Y|x)) = \sum_y P_\theta(y|x) \log \frac{P_\theta(y|x)}{P_{\mathcal{C}}(y|x)} \tag{2.8}$$

2.4.3　主动学习在其它机器学习任务中的应用

在其它基于监督学习的自然语言处理任务中，主动学习被广泛用来解决减少数据标注工作量的问题。Chen 等人[39] 将主动学习用于基于 SVM 的命名实体识别 (Named Entity Recognition) 任务，在保证模型性能的前提下，相比基线方法降低了 42% 的标注量。Cormack 等人[40] 针对文本分类 (Text Classification) 任务，对主动学习在不同监督学习模型上的效果做了相关研究。实验结果表明在不同的监督学习模型上，文本分类需要的训练样本数量都有所减少。

与已有的主动学习方法不同，Alonsod 等人[41] 为了避免主动学习过程中选择到离群点，提出了一种基于概率的样本选择方法。该方法并非在每轮迭代都选择不确定度最高的样本，而是根据当前样本分类结果的不确定度，对样本赋予不同的被选择的概率。不确定度越高，样本被选中的概率越大，反之，被选中的概率越低。该方法在词义标注 (Word Sense Annotation) 任务中取得了明显的效果。

在语料库辅助标注任务中，Ayache 等人[42] 为了给 TRECVID 2007 提供视频语料，开发了基于 Web 的标注工具。由于时间和人力资源有限，Ayache 等人利用主动学习方法，通过算法选择信息量较大的样本进行标注，并通过已标注样本影响未标注样本的预测结果，提高人工标注效率。实验表明，通过主动学习方法，仅需要人工标注 30% 的语料，得到的银标准语料对模型训练任务就能达到和金标准语料相同的性能。

2.4.4　已有方法和本文工作的联系与区别

主动学习在多种机器学习任务中已有较多的研究成果，但目前在实体链接任务中还没有关于用主动学方法降低训练样本数量的研究。因此，本文将主动学习方法用于实体链接任务，来降低人工标注工作量。同时，针对实体链接任务的特点，本文对样本选择方法做了改进。最后本文通过实验证明，与非主动学习的标注方法相比，基于主动学习方法的实体链接方法极大地降低了人工标注工作量，并且，经过改进的主动学习策略能显著提升主动学习器的性能。

另外，主动学习方法目前主要应用于监督学习模型的训练，然而对于银标准语料的辅助构建，主动学习应用较少。在实体链接任务中，考虑到成本有限，应尽可能提高标

注语料的质量。因此，本文借助主动学习方法对实体链接语料库辅助标注做了相应研究。

2.5 本章小结

本章介绍了实体链接任务的相关研究。首先简单介绍了实体链接任务的概念，并对目前用于处理实体链接任务的模型和方法进行了相关介绍。然后介绍了已有的几种主动学习方法，并对主动学习方法在其它机器学习任务中的应用进行了相关介绍。最后将已有方法和本文的工作做了对比与分析，给出了联系与区别。

第三章　基于主动学习的实体链接模型训练

本章以监督学习模型处理实体链接任务，研究主动学习方法在减少标注训练样本数量上的作用。本章首先介绍基于监督学习的实体链接方法，包括模型使用的特征以及模型对实体链接任务的处理方法。接着，本章介绍如何将主动学习方法应用于实体链接任务，并对初始样本选择方法和迭代样本选择方法做了改进。最后给出了实验结果和实验分析。

3.1　基于监督学习的实体链接方法

用监督学习方法处理实体链接任务主要分为两个步骤，第一步是特征提取，第二步是利用模型对输入特征向量进行处理，并输出预测的目标实体。

3.1.1　实体链接特征

在给定指称项 m 和指称项对应的候选实体集 $E_m = \{e_1, e_2, ..., e_n\}$ 后，计算模型的输入特征:

(1) 候选实体 e_i 在知识库中的先验概率，$PriorInKB(m, e_i)$，该特征的定义如公式3.1所示。

$$PriorInKB(m, e_i) = \frac{count(e_i)}{\sum_{e_i \in E_m} count(e_k)} \tag{3.1}$$

其中，$count(e)$ 表示实体 e 在知识库中作为锚文本出现的次数。

(2) 指称项上下文与候选实体对应词条摘要的文本相似度，$ContextSimilarity(m, e_i)$，该特征的定义如公式3.2所示。

$$ContextSimilarity(m, e_i) = \frac{coocurence(m, e_i)}{length(m)} \tag{3.2}$$

其中，$coocurence(m, e_i)$ 表示指称项 m 上下文和候选实体 e_i 对应知识库中摘要文本相同的单词数，$length(m)$ 表示指称项 m 的上下文长度。

(3) 带标注样本集中，候选实体 e_i 的流行度，$PopInCorpus(e_i)$，该特征会随着人工标注的进行而产生变化。该特征的定义如公式3.3所示。

$$PopInCorpus(e_i) = \frac{\sum_{anno \in Corpus} I(anno.e, e_i)}{sizeof(Corpus)} \tag{3.3}$$

其中，$anno$ 表示已标注语料中的一条标注记录，$anno.e$ 表示标注记录的目标实体。$I(e_i, e_j)$ 是示值函数，实体 e_i 和实体 e_j 相同则为 1，否则为 0。$sizeof(Corpus)$ 表示已标注指称项个数。

(4) 带标注样本集中，候选实体 e_i 作为指称项 m 的目标实体的先验概率，$Prior(m, e_i)$，同上，该特征也会随着标注的进行而产生变化。该特征的定义如公式3.4所示。

$$Prior(m, e_i) = \frac{\sum_{anno \in Corpus} I(anno.e, e_i) \times I(anno.m, m)}{\sum_{anno \in Corpus} I(anno.m, m)} \tag{3.4}$$

其中，$anno.e$ 的定义同上。$anno.m$ 表示标注记录 $anno$ 的指称项。

(5) 指称项名字和候选实体标题的编辑距离，$EDSimarity(m, e_i)$。该特征的定义如公式3.5所示。

$$EDSimarity(m, e_i) = \begin{cases} 1 & |length(m) - length(e_i)| = ED(m, e_i) \\ 0 & \text{其它} \end{cases} \tag{3.5}$$

编辑距离 $ED(\cdot, \cdot)$ 是指两个字符串进行转换时，需要的字符级最少操作次数。编辑距离相似度能够检测指称项和候选实体之间的别名、缩略名等关系。

3.1.2 目标实体预测

在使用监督学习模型 C 处理实体链接任务时，首先需要训练模型 C。训练样本由带标注的指称项集合构成，例如，带标注的指称 m 包含 n 个候选实体，指称 m 的目标实体是 e^*，任意一个候选实体 e_i 和指称项 m 组成一个样本对 $\langle m, e_i \rangle$，若 e_i 是目标实体 e^*，样本对标记为正例，否则标记为反例。用所有带标注的样本对训练得到模型 C。

训练得到模型 C 后，对于未标注指称 m 及其候选实体集 E_m，根据上一节的计算方法获得所有候选实体和指称项组成的样本对的特征向量，模型 C 根据输入的特征向量，计算每个候选实体是目标实体的概率 $P_C(e_i|m)$。该预测概率的计算方式和模型的选择相关，例如选择支持向量机模型，可以用样本点距离分类超平面的距离估计正确分类的概率[43]；选择人工神经网络模型，可以对输出层每个节点的权值做 $softmax$ 处理，以此获得每个分类标签的分类预测概率[44]，在此不再一一例举其它模型的预测概率估计方法。取概率最高的候选实体作为指称项 m 的预测实体，如公式3.6所示。

$$\hat{e} = \underset{e_i \in E_m}{\text{argmax}} P_C(e_i|m) \tag{3.6}$$

然后将该预测实体被正确划分的概率作为指称 m 被正确链接的概率，如公式3.7所示。

$$P_C(\hat{e}|m) = \max_{e_i \in E_m} P_C(e_i|m) \tag{3.7}$$

3.2　实体链接的主动学习算法

为了减少人工标注的工作量，本文采用基于池的主动学习方法 (Pool-based Active Learning) 选择实体链接的待标注指称项。主动学习流程图如图3.1所示。

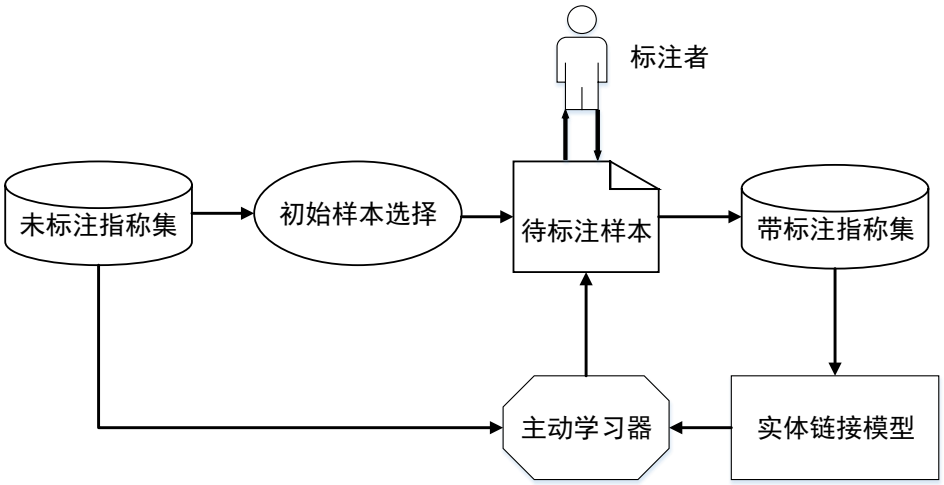


图 3.1: 主动学习流程概览

在每轮迭代训练中，主动学习器能够选出信息量最大的未标注样本集交由人工标注。已有的主动学习[45]流程如算法3.1所示。在给定未标注训练样本集后，主动学习进程分为两个阶段，第一阶段的工作是选择初始训练样本集并以此训练初始分类器，第二阶段的工作是选择最佳标注样本并以此迭代训练模型。

算法第1-2 行对应主动学习进程的第一个阶段，已有的做法是在未标注样本集 $\mathcal{U}$ 中随机选择待标注样本 $\mathcal{U}_{selected}$，然后对这些未标注样本进行人工标注，通过这种方式得到初始带标注样本集 $\mathcal{L}_0$，然后用 $\mathcal{L}_0$ 训练得到初始实体链接模型 C。由于当初始带标注样本集合较小时，通过随机选择的方式获得的子样本集可能无法很好地代表整个样本集。因此，本文对初始训练样本选择方法做了改进，提出基于指称项流行度的初始样本选择方法。

算法第3-9 行对应主动学习的第二个阶段，其中对于算法第4行信息量度量方法的选择，已有的做法是将样本分类结果的不确定度大小作为信息量的衡量标准，第 t 轮迭代中选出的样本会被加入到带标注样本集 $\mathcal{L}_t$ 中，然后用 $\mathcal{L}_t$ 重新训练模型。在实体链接任务中，仅将不确定度作为选择待标注样本的依据，可能会导致选择过多的离群样本点，

算法 **3.1** 基于主动学习的实体链接任务训练进程

输入**:** 未标注的指称样本集 $\mathcal{U} = \{m^{(u)}\}_{u=1}^{U}$

输出**:** 实体链接分类器 C

从未标注训练集 $\mathcal{U}$ 中选择并标注初始训练样本集 $\mathcal{L}_0$

利用初始训练样本集 $\mathcal{L}_0$ 训练得到弱分类器 $C = train(\mathcal{L}_0)$

repeat

  从 $\mathcal{U}$ 中找出 k 个信息量最大的样本组成样本集 $\mathcal{U}_{selected} = \{m^{(u)}\}_{u=1}^{k}$

  对 $\mathcal{U}_{selected}$ 中的指称进行人工标注得到 $\mathcal{L}_{selected} = \{\langle m, e\rangle^{(l)}\}_{l=1}^{k}$

  $\mathcal{U} = \mathcal{U} \setminus \mathcal{U}_{selected}$

  $\mathcal{L}_{t+1} = \mathcal{L}_t \cup \mathcal{L}_{selected}$

  $C = train(\mathcal{L}_{t+1})$

until 达到预期精度或样本集已全部标注

return C

不利于模型的训练。因此，本文对迭代训练样本选择方法做了改进，提出综合分类不确定度和指称项流行度的迭代训练样本选择方法。

3.3 改进的初始训练样本选择方法

本节首先对改进的初始训练样本选择方法进行了详细说明，然后给出了该方法对应的算法描述。

3.3.1 初始训练样本选择方法

在主动学习的初始阶段，需要产生一个初始样本集用于训练初始模型。已有的主动学习方法采用随机选择的方式产生初始训练样本集，由于训练样本集相对较小，随机选择的方式很难保证初始样本集的代表性。但是，训练一个性能较好的初始模型对提高主动学习收敛速度非常重要。因此，本文对已有的基于随机选择的初始样本集生成方法做了改进。

为了提高实体链接初始模型的性能，本文提出基于流行度的初始样本选择方法(Sampling by Popularity, SBP)。指称项的流行度按照相同名字的指称项在语料库中出现的频率计算。例如，语料库中包含 n 个指称项，名字为“Jordan”的指称项出现了 m 次，则名字为“Jordan”的所有指称项的流行度为 m/n。该样本选择方法首先对指称项按照指称项名字分类，计算指称项流行度，然后对指称项流行度高的样本做标注，加入初始训练样本集。该方法的目的是在初始样本选择阶段，尽量选择出现频率较高的指称项，以此保证初始样本集的代表性，从而提高在初始样本集较小的情况下，尽可能提高初始模型的性能。

3.3.2　初始训练样本选择算法

算法3.2是对基于流行度的初始样本选择方法的算法描述。初始训练样本的选择分为两个步骤，第一步是不同指称项流行度的统计；第二步是根据指称项流行度选择初始训练样本。

算法第1-2行对未标注样本集按指称项名字进行分类，例如所有指称名字为“Jordan”的训练样本都会被划分到集合 Ψ_{Jordan} 中。然后根据集合中的样本数量计算各个指称项在训练集中的流行度，指称项名字相同的样本具有相同的流行度。

算法第3-7行对指称项流行度做排序，并在流行度最高的 k 个样本集中分别随机选择一个或多个[1]样本，经过人工标注后加入初始训练样本集。这样做的目的在于避免初始训练样本集中的样本指称项名字重复度过高，同时保证尽量选择流行度高的指称项样本，从而提高初始训练样本集的代表性。

算法 3.2 基于流行度的初始样本集选择算法

输入： 未标注的实体链接样本集 $\mathcal{U} = \{m^{(u)}\}_{u=1}^{U}$，初始训练集样本个数 k

输出： 初始样本集 $\mathcal{L}$

1: 对未标注训练样本按照指称名字 $name$ 分组，指称名字为 $name$ 的样本被分到集合 Ψ_{name} 中
2: 计算名字为 $name$ 的指称的流行度 $\Phi_{name} = \frac{|(\Psi_{name})|}{U}$
3: 对不同指称的流行度排序，取流行度最高的 k 个指称的名字 $\mathcal{S} = \{name_1, name_2, ..., name_k\}$
4: **for** each $name$ in $\mathcal{S}$ **do**
5: 　从集合 Ψ_{name} 中随机取一个或多个样本并对其人工标注得到带标注的样本对 $L = \langle m, e \rangle$
6: 　$\mathcal{L} = \mathcal{L} \cup \{L\}$
7: **end for**
8: **return** $\mathcal{L}$

3.4　改进的迭代训练样本选择方法

本节首先对改进的迭代训练样本选择方法进行了详细说明，然后给出了该方法对应的算法描述。

3.4.1　迭代训练样本选择方法

在初始训练集上训练得到初始模型后，需要迭代选择信息量最大的未标注样本做标注，然后对模型做重新训练。在已有的主动学习迭代训练样本选择过程中，通常将不确

[1]当指称名字数量小于 k 时，需要在一个类簇中选择多个样本，保证最终获得 k 个样本。

定度最大的样本作为信息量最大的样本，这些样本会交由人工标注并加入下一轮迭代训练，通过这种方式迭代逼近真实模型。本文采用间隔（Margin）度量未标注样本的不确定度。

如公式3.8所示，e^{*1} 和 e^{*2} 分别表示在当前模型 C 中，给定指称项 m，对应置信度最高的和次高的候选实体，以这两个候选实体的置信度差的绝对值作为指称 m 链接到正确实体的置信度。

$$Confidence(m) = P_C(e^{*1}) - P_C(e^{*2})$$

其中，

$$\begin{aligned} e^{*1} &= \underset{E_i \in E_m}{\text{argmax}}\, P_C(e_i|m) \\ e^{*2} &= \underset{E_i \in E_m \setminus \{e^{*1}\}}{\text{argmax}}\, P_C(e_i|m) \end{aligned} \tag{3.8}$$

易知，$Confidence(m)$ 越小，表示模型 C 越难正确地从候选实体 e^{*1} 和候选实体 e^{*2} 中区分出目标实体，因此，指称项 m 被正确链接的置信度越小，指称项被正确链接的不确定度越大。指称项链接的不确定度如公式 (3.9) 所示。

$$Uncertainty(m) = 1 - Confidence(m) \tag{3.9}$$

在本文的实体链接任务中，如果采用仅基于不确定度的样本选择方法来选择待标注样本，则会存在两个缺陷。第一，存在一些不确定度较大，但是在所有样本中处于边缘位置的样本，这些离群样本点不具有代表性，相反还可能会对模型的训练产生负作用，在样本选择过程中，应该尽量避免这类样本点；第二，本文实体链接模型的输入特征向量包含已标注语料中候选实体流行度（$PopInCopus(e_i)$）和已标注语料中候选实体先验概率（$Prior(e_i|m)$）这两个特征维度，这两个特征都是在已标注样本集上通过统计得到的，因此，带标注的样本分布越离散，带标注样本集中包含的指称项和实体越多，那么上述两个特征维度的统计量越具有代表性。从这个角度看基于不确定度的样本选择算法对这两个特征的计算可能会产生负面影响。因此主动学习的样本选择策略需要在不确定度和多样性之间取一个权衡。为了解决这个问题，本文提出综合不确定度和流行度的样本选择方法（Sampling by Uncertainty and Popularity, SUP）。

3.4.2 综合不确定度和流行度的样本选择算法

接下来给出综合不确定度和流行度的样本选择方法的算法描述。SUP 算法是对不确定度和多样性的折衷，如算法3.3所示。

算法第2行将所有未标注样本集中的指称项按照指称项流行度进行聚类，完成聚类后，同一个类簇的样本流行度相近。

算法 **3.3** 综合不确定度和流行度样本集选择算法

输入：　未标注的实体链接样本集 $\mathcal{U} = \{m^{(u)}\}_{u=1}^{U}$
　　选择标注的样本个数 k

输出：　本轮选择标注的训练样本集 $\mathcal{L}$

1: $\mathcal{L} = \{\}$
2: 用 K-means 聚类法将未标注训练样本集按照指称的流行度划分为 k 类，$Clusters = \{\Psi_1, \Psi_2, ..., \Psi_k\}$
3: **for** each Ψ_i in $Clusters$ **do**
4: 　$m^* = \text{argmax}_{m \in \Psi_i} Uncertainty(m)$
5: 　对 m^* 人工标注，得到 $\langle m^*, e \rangle$
6: 　$\mathcal{L} = \mathcal{L} \cup \{\langle m, e \rangle\}$
7: **end for**
8: **return** $\mathcal{L}$

算法第4行从各个流行度的类簇中分别选出不确定度最大的样本交由人工标注，这里不确定度的计算方式依然是模型预测的最优候选实体和次优候选实体置信度的间隔。该样本选择算法能够保证各个指称项流行度区间的样本都有均等机会被选中标注，同时又能保证在每个类簇中被选中标注的样本都是这个类簇中不确定度最大的样本，在兼顾所选样本集多样性的同时保证了不确定度。

3.5　实验结果与分析

本节介绍主动学习方法在实体链接模型训练中的实验，说明实验环境、实验所用数据集、评价指标以及实验配置，最后给出了实验结果数据，并对实验结果进行了分析。

3.5.1　实验环境

硬件环境：

- CPU: Intel(R) Core(TM) i7-3770 CPU @ 3.40GHz
- Memory: 4×8GB DDR3 1333 ECC

软件环境：

- CentOS Linux release 7.3.1611 (Core)
- IntelliJ IDEA 2016.3
- Java 1.8

- MySQL 5.6

开源工具：

- LibSVM[2]
- Weka 3.6[3]
- Stanford NLP[4]

3.5.2 实验数据集

本文使用的数据集是在 *CoNLL'03* 的基础上做了实体链接标注的 *Aida* 数据集。数据集包含 1393 篇文章和 28815 个可用指称项。本文取 20000 个指称项作为训练集，其余的指称项作为测试集。

3.5.3 评价指标

本章研究主动学习在实体链接模型训练中的作用，并验证本文提出的基于流行度的初始样本选择方法以及综合不确定度和流行度的迭代训练样本选择方法对减少人工标注样本工作量的效果。实际上数据集中的指称项都已经标注好对应的目标实体，但是在模型训练过程中，本文不直接使用 Aida 数据集的标注结果，只有通过主动学习方法选择的样本的标注结果才会被使用，以此来模拟人工标注的过程。本章实验分为以下两个方面:

(1) 以随机选择方法为基线方法，评价基于流行度的初始样本选择方法在标注相同数量样本的前下，对初始模型性能提升的效果。

(2) 以随机选择法为基线方法，评价基于不确定度的样本选择方法以及综合不确定度和流行度的样本选择方法对模型训练收敛速度的提升效果。

本文对各种样本选择方法好坏的评价指标为主动学习过程中，由主动学习器选择的训练样本所训练得到的模型在测试集里的性能表现，性能表现体现在测试集上实体链接的正确率，计算方法如公式3.10所示。

$$准确率(P) = \frac{测试集中被正确链接的指称项个数}{测试集中所有指称项个数} \times 100\% \tag{3.10}$$

[2]用于实现支持向量机。

[3]用于配合 LibSVM 实现支持向量机训练。

[4]用于实现英文分词及词干抽取。

3.5.4　模型选择

本章使用监督学习模型处理实体链接任务，并将该任务作为分类问题来处理。在常用的机器学习分类器中，SVM 模型[46] 是目前分类效果较好的模型，其可以通过核函数（本文使用径向基函数（Radial Base Function, RBF））将非线性可分的输入空间的特征映射到高维线性可分的空间的特征，以此学习非线性可分模型。此外，样本点距离分类超平面的远近可以作为分类置信度的依据，而主动学习过程中，分类置信度恰好是样本选择策略的考虑因素之一。

综上，本文在实验阶段使用 SVM 模型作为监督学习模型的代表，处理实体链接任务并研究主动学习方法在该任务上发挥的作用。

3.5.5　初始训练样本选择方法实验

为了验证基于指称项流行度的初始训练样本选择方法对提高初始模型性能的效果，在实验环节，分别通过基于随机选择方法 (Random) 和基于流行度选择方法 (SBP) 得到初始样本集，对比在不同初始训练样本集大小的情况下，由初始样本集训练得到的初始模型的性能。

实验中，初始样本集大小的设置从 2000 开始，依次以 2000 递增，最大的初始样本集包含 18000 个训练样本。从图3.2可以看出，在初始样本集较小的时候，基于指称项流行度的样本选择方法比随机样本选择方法在初始模型的性能上有显著的提升，在初始样本集大小为 2000 的时候，性能提升了 10.1%，在初始样本集大小为 4000 的时候，性能提升了 5.3%。当初始样本集大小继续增加时，性能提升值逐渐减小，甚至在一些情况下提升值为负。总体来看，当初始样本集较大时，两种初始训练样本选择方法性能差别不大。

该实验表明，在初始样本集较小的时候，由于样本分布的不均匀特性，随机选择的方法很难选择出能代表整个数据集的样本子集。在实际应用中，用于训练初始模型的初始训练样本集通常较小，使用基于指称项流行度的初始样本选择方法，能有效提升初始模型性能，加快模型在后续训练过程的收敛速度。

3.5.6　迭代训练样本选择的实验

为了验证主动学习方法中不同的样本选择策略对实体链接模型训练的收敛速度的影响，本文在实验环节，对五种不同组合的样本选择策略进行了比较。五种组合如下所示:

(1) Random，初始训练样本和后续迭代训练样本选择都采用随机选择方法，作为基线方法。

(2) Random+Uncertainty，初始训练样本采用随机选择方法，后续迭代训练样本采用基于不确定度的选择方法。

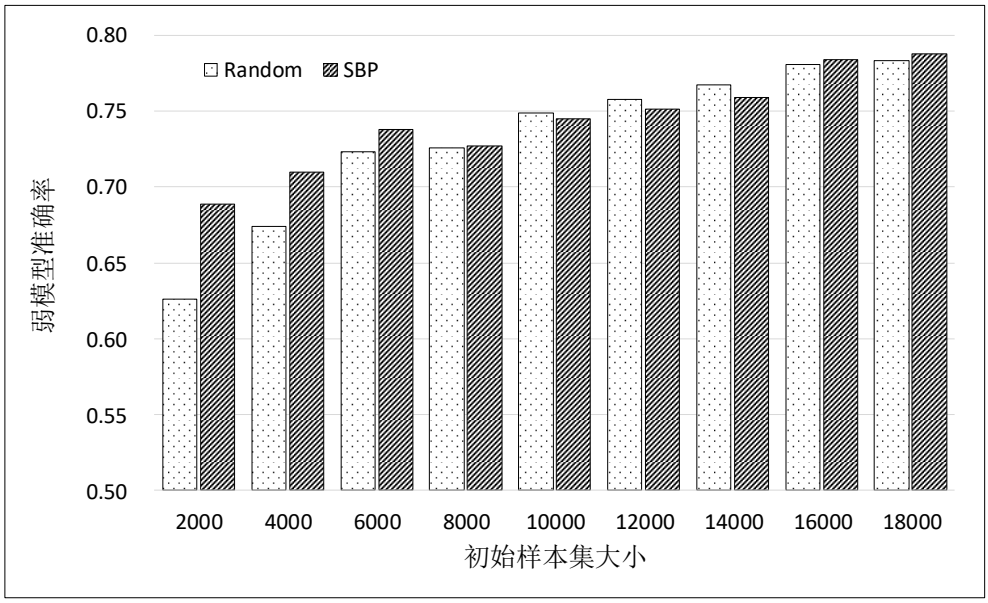


图 3.2: 初始训练样本选择实验结果

(3) Random+SUP，初始样本采用随机选择方法，后续迭代训练样本采用综合不确定度和流行度的选择方法。

(4) SBP+Uncertainty，初始样本采用基于流行度的选择方法，后续迭代训练样本采用基于不确定度的选择方法。

(5) SBP+SUP，初始样本采用基于流行度的选择方法，后续迭代训练样本采用综合不确定度和流行度的选择方法。

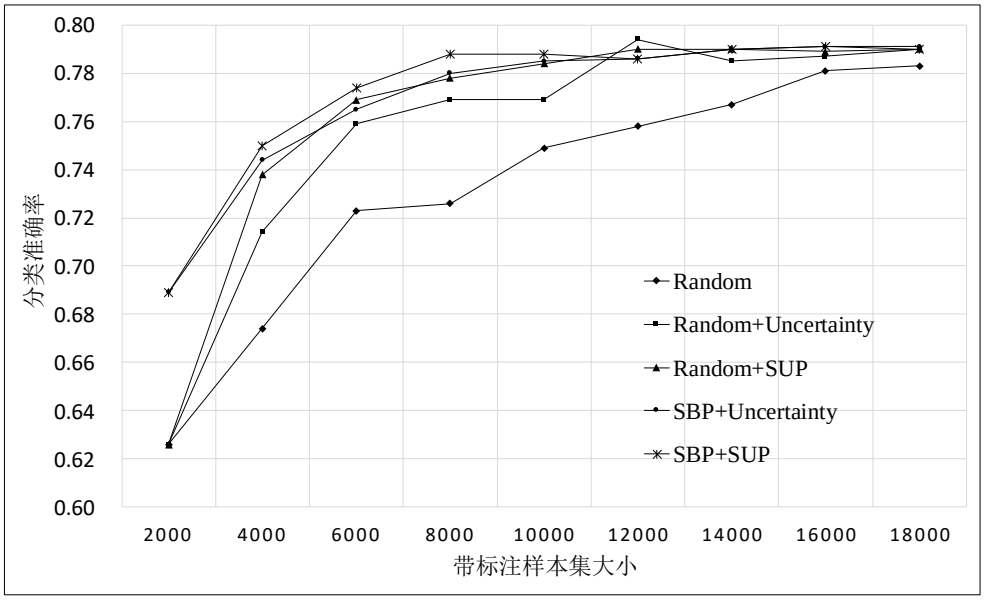


图 3.3: 迭代训练样本选择实验结果

从图3.3展示的实验结果可以看出，相比随机选择的基线方法，其它四种基于主动

学习的样本选择策略对加快模型训练的收敛速度都有显著的提高。证明主动学习方法在实体链接任务中，对减少人工标注样本的工作量确实是有显著效果的。另外，观察曲线走势，综合不确定度和流行度的样本选择方法要优于基于不确定度的样本选择方法，这也验证了基于不确定度的样本选择方法并不能找到最具信息量的样本集，主动学习样本选择过程中需要兼顾不确定度和多样性。

为了量化各种主动学习样本选择策略的表现，采用基于 $deficiency$ 值[47] 的度量方式进行评价，如公式3.11所示。

$$Def_n(AL, REF) = \frac{\sum_{t=1}^{n}(acc_n(REF) - acc_t(AL))}{\sum_{t=1}^{n}(acc_n(REF) - acc_t(REF))} \tag{3.11}$$

其中，$acc_n(REF)$ 和 $acc_n(AL)$ 分别表示第 t 轮迭代训练基线方法和主动学习方法训练的模型在测试集上的准确率。主动学习器性能越好，等式分母中 $acc_t(AL)$ 值越大，则等式计算值越小。因此，$Def_n(AL, REF)$ 越小，主动学习效果越好。

表 3.1: 主动学习策略结果评价

Random	Random+Uncertainty	Random+SUP	SBP+Uncertainty	SBP+SUP
1.000	0.552	0.369	0.274	**0.190**

表3.1是对主动学习效果的定量分析，无论基于随机的初始样本选择还是基于流行度的初始样本选择，在后续迭代训练中采用基于不确定度和流行度的样本选择算法都能比基于不确定度的样本选择算法得到更快的训练收敛速度。并且采用基于流行度的初始训练样本选择配合综合不确定度和流行度的迭代训练样本选择的策略，$deficiency$ 值是最低的，性能表现最优。

3.6　本章小结

本章针对实体链接问题，基于主动学习方法，研究了减小人工标注工作量的方法。并且在已有主动学习方法的基础上，针对实体链接任务提出了基于流行度的初始样本选择算法、综合不确定度和流行度的迭代训练样本选择算法。实验结果表明，主动学习方法可以有效地降低实体链接训练集的标注数据，并保持较高的泛化能力。在初始样本集生成阶段，本文提出的选择算法相比随机选择的基线算法初始分类器性能升了10.1%。在主动学习迭代训练阶段，本文提出的选择算法相比基于不确定度的选择算法 $deficiency$ 值降低了 16.1%。

第四章　基于主动学习的实体链接语料构建

本章以无监督学习模型辅助实体链接银标准语料的构建，研究主动学习方法在减少人工标注语料数量上的作用。首先介绍基于图的无监督学习的实体链接方法，包括模型使用的特征以及模型对实体链接任务的处理方法。接着，介绍如何使用主动学习方法选择需要人工辅助标注的样本。最后给出实验结果和实验分析。

4.1　基于图的协同推断

构建语料的常用方法是借助已有模型进行辅助构建，本节将首先给出用于辅助标注实体链接语料的基于图的协同推断方法，这是 Han 等人[18] 提出的一种基于无监督学习的实体链接方法。

在例子“Michael Jordan, the Chicago Bulls retired professional basketball player, starred in the 1996 feature film Space Jam.”中，已有实体链接方法会通过词袋模型计算指称项“Michael Jordan”和目标实体 *Michael Jordan* 的相似度，例如实体 *Michael Jordan* 的维基百科页面摘要中包含“Jordan played 15 seasons in the National Basketball Association (NBA)”，可以看出，由于指称项上下文和实体摘要文本的单词完全不相同，该种方式计算所得的相似度为 0。但是“Chicago Bulls”和“NBA”之间却是具有很强的语义关联度的。为了解决这种语义的问题，基于图的协同推断方法借助知识库中的相关信息，建立了实体之间的相关联系，在考虑单个指称项与其对应的候选实体之间的相似度的同时，还考虑了在同一个文档中，不同的指称项之间的候选实体的关联度。

4.1.1　实体-实体相似度

同一篇文档中的实体之间一般会有相互关联。在处理实体链接任务时，需要通过定量的方式计算两个实体之间的相似度。张涛等人[48] 通过知识库（维基百科）中的链接关系计算实体间的相似度。计算过程中考虑到了其他指称项对应的实体，因而也称为全局相似度。

图4.1是实体 *Michael Jordan* 在维基百科词条中的部分文本。可以看到，词条文本包含了大量的指向其它实体的链接，例如 National Basketball Association、Chicago Bulls、Washington Wizards 等。本文通过统计维基百科实体页中的实体链接情况，统计出各个实体在其他实体对应的实体页中被链接的次数，然后计算实体实体之间的相似度，例如，实体 a 和实体 b 之间的相似度可以表示为：

$$SR(a,b) = 1 - \frac{\log(\max(|A|,|B|)) - \log(|A \cap B|)}{\log(|W|) - \log(\min(|A|,|B|))} \tag{4.1}$$

Michael Jordan

From Wikipedia, the free encyclopedia

For other people named Michael Jordan, see Michael Jordan (disambiguation).

Michael Jeffrey Jordan (born February 17, 1963), also known by his initials, **MJ**,[3] is an American retired professional basketball player, businessman, and principal owner and chairman of the Charlotte Hornets. Jordan played 15 seasons in the National Basketball Association (NBA) for the Chicago Bulls and Washington Wizards. His biography on the NBA website states: "By acclamation, Michael Jordan is the greatest basketball player of all time."[4] Jordan was one of the most effectively marketed athletes of his generation and was considered instrumental in popularizing the NBA around the world in the 1980s and 1990s.[5]

Jordan played three seasons for coach Dean Smith at the University of North Carolina. As a freshman, he was a member of the Tar Heels' national championship team in 1982. Jordan joined the Bulls in 1984 as the third overall draft pick. He quickly emerged as a league star, entertaining crowds with his prolific scoring. His leaping ability, demonstrated by performing slam dunks from the free throw line in slam dunk contests, earned him the nicknames **Air Jordan** and **His Airness**. He also gained a reputation for being one of the best defensive players in basketball.[6]

图 4.1: 维基百科中 *Michael Jordan* 的页面

公式4.1中，A 和 B 分别表示维基百科中对应实体页中包含链接到实体 a 和实体 b 的实体集合。W 表示维基百科中所有实体的集合。在示例文本中，包含“Michael Jordan”、“Chicago Bulls”、“Space Jam”这三个指称项，其中“Michael Jordan”指向的实体可能是 *Michael Jordan*，也可能是 *Michael B. Jordan*。计算这两个实体和其它两个指称项对应的实体之间的相似度，结果如表4.1所示。

表 4.1: 实体间相似度例子

	Space Jam	Chicago
Michael Jordan	0.66	0.82
Michael B. Jordan	0.00	0.00

观察该表可以发现，当确定文本中指称项“Chicago Bulls”和“Space Jam”所指向的实体分别是 *Chicago Bulls* 和 *Space Jam* 之后，计算 *Michael Jordan* 与 *Chicago Bulls*、*Space Jam* 的实体相似度分别是 0.66 和 0.82，*Michael B. Jordan* 与 *Chicago Bulls*、*Space Jam* 的实体相似度都是 0.00。由此可以推断，*Michael Jordan* 与上下文环境的实体相似度更大，更有可能是指称项“Michael Jordan”对应的目标实体。

4.1.2　指称项-实体相似度

指称项-实体相似度是指在给定指称项 m 和候选实体 e 后，根据指称项 m 上下文等特征计算得到的和候选实体 e 之间的相似度。计算相似度时，因为只考虑单个指称项，因而也称为局部相似度。

计算指称项-实体相似度需要为指称项的上下文定义一个窗口大小，通过计算该窗口文本和候选实体词条摘要文本的相似度得到指称项-实体相似度。本文参考 Pedersen

等人[49] 的相关研究，将上下文的窗口大小定义为 50。

本文采用词袋模型计算指称项-实体相似度，给定指称项 m 和候选实体 e，指称项-实体相似度如公式4.2所示：

$$CP(m,e) = \frac{m \cdot e}{|m||e|} \tag{4.2}$$

其中，m 和 e 分别是指称项窗口文本和候选实体知识库摘要文本的词向量表示，向量中的每个维度分别是对应单词的 TFIDF 表示。

4.1.3 构建协同推断图

协同推断图的每个节点可以是指称项，也可以是指称项对应的候选实体。图中的边分为两种类型，第一类是指称项节点和候选实体节点之间的边，这类边的权值用指称项-实体相似度表示。第二类是候选实体节点和其它候选实体节点之间的边，这类边的权值用实体-实体相似度表示。由于一个指称项对应的候选实体集包含的候选实体数量可能非常大，那么协同推断图的节点和边的数量可能也会非常大。图的复杂度太高会直接影响后续在图上相关计算的效率，为了解决这个问题，本文对连接边权值较小的候选实体节点和边做了删除处理。

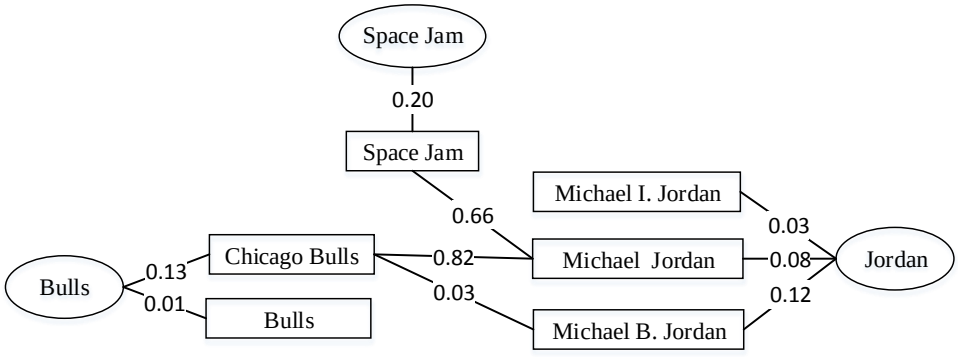


图 4.2: 协同推断图示例图

图4.2是实体链接任务中协同推断图的一个例子。图中椭圆边框的节点表示指称项，矩形边框的节点表示指称项对应的候选实体。例如指称项“Jordan”的候选实体包含 *Michael I. Jordan*、*Michael Jordan*、*Michael B. Jordan*。指称项“Jordan”与其对应的候选实体之间的边的权值是指称项和实体之间的局部相似度。不同指称项对应的候选实体之间的边为全局相似度，例如指称项“Jordan”对应的候选实体 *Michael Jordan* 和指称项“Bulls”的候选实体 *Chicago Bulls* 的实体相似度为 0.82。图中并非所有节点都有连接边，这是因为一些边的权值很小，为提高后续计算效率而被删除。

4.1.4　协同推断方法

构建好协同推断图以后，通过随机行走（Random Walk, RW）算法[32] 对候选实体的置信度进行推断，选择置信度最高的候选实体作为预测目标实体。

协同推断图中包含 n 个节点（k 个指称项节点和 l 个候选实体节点）。首先需要计算文本中所有指称项在上下文中的重要程度，如公式4.3所示。

$$Importance(m) = \frac{tfidf(m)}{\sum_{m \in D} tfidf(m)} \tag{4.3}$$

初始证据向量 s 的维度为 $k+l$，对 s 的初始化方法如公式4.4所示。

$$s_i = \begin{cases} Importance(m) & \text{节点}i\text{对应指称项}m \\ 0 & \text{节点}i\text{对应候选实体} \end{cases} \tag{4.4}$$

证据传递矩阵 T 是一个大小为 $(k+l) \times (k+l)$ 的矩阵。该矩阵是一个稀疏矩阵，定义如公式4.5所示。公式中 E_m 表示指称项 m 对应的候选实体集，N_e 表示推断图中与实体 e 有边相连的实体集合。

$$T_{ij} = \begin{cases} \frac{CP(m,e)}{\sum_{e \in E_m} CP(m,e)} & \text{节点}j\text{对应指称项}m\text{，节点}i\text{对应候选实体}e \\ \frac{SR(e_i,e_j)}{\sum_{e \in N_{e_j}} SR(e_i,e)} & \text{节点}j\text{对应候选实体}e_j\text{，节点}i\text{对应候选实体}e_i \\ 0 & \text{其它} \end{cases} \tag{4.5}$$

在实体链接任务中，随机行走算法如算法4.1所示。

算法 4.1 基于图的协同推断中的随机行走算法

输入： 初始证据向量 s，证据传递矩阵 T

输出： 图的稳定状态证据向量 r^*

1: 初始化 $r^0 = s$
2: **repeat**
3: $r^{t+1} = (1-\lambda) \times T \times r^t + \lambda \times s$
4: **until** r 得到稳定状态或者迭代次数达到上限
5: **return** r^*

由于 T 是一个稀疏矩阵，协同推断图中存在一些候选实体节点没有出边，因此算法4.1的第3行，在随机行走的迭代过程中，加入了证据重分配率参数 λ[1]，这是为了在随机行走过程中以一定概率从起始点重新进行随机行走，解决某些节点无法进行证据传播的问题。参考 Hu 等人[50] 的相关研究，算法第4行的迭代次数上限，本文选取的值为100。

[1]本文通过实验，取 $\lambda = 0.5$，效果最佳。

以图4.2为例，初始证据向量 $s=(0.41,0.33,0.26,0,0,0,0,0,0)^T$，$r^0=s$。随机行走过程第 1 次迭代如公式4.6所示。

$$\begin{pmatrix}0.21\\0.17\\0.13\\0.03\\0.07\\0.11\\0.15\\0.01\\0.13\end{pmatrix}=.5\times\begin{pmatrix}0&0&0&0&0&0&0&0&0\\0&0&0&0&0&0&0&0&0\\0&0&0&0&0&0&0&0&0\\0.13&0&0&0&0&0&0&0&0\\0.35&0&0&0&0&0&0.96&0&0\\0.52&0&0&0&0&0&0.04&0&0\\0&0.92&0&0&0.55&1&0&0&0\\0&0.08&0&0&0&0&0&0&0\\0&0&1&0&0.45&0&0&0&0\end{pmatrix}\times\begin{pmatrix}0.41\\0.33\\0.26\\0\\0\\0\\0\\0\\0\end{pmatrix}+.5\times\begin{pmatrix}0.41\\0.33\\0.26\\0\\0\\0\\0\\0\\0\end{pmatrix}\tag{4.6}$$

在完成第 100 轮随机行走迭代过程后，达到了稳定状态下，最后一轮迭代如公式4.7所示。

$$\begin{pmatrix}0.21\\0.17\\0.13\\0.01\\0.10\\0.06\\0.13\\0.01\\0.09\end{pmatrix}=.5\times\begin{pmatrix}0&0&0&0&0&0&0&0&0\\0&0&0&0&0&0&0&0&0\\0&0&0&0&0&0&0&0&0\\0.13&0&0&0&0&0&0&0&0\\0.35&0&0&0&0&0&0.96&0&0\\0.52&0&0&0&0&0&0.04&0&0\\0&0.92&0&0&0.55&1&0&0&0\\0&0.08&0&0&0&0&0&0&0\\0&0&1&0&0.45&0&0&0&0\end{pmatrix}\times\begin{pmatrix}0.21\\0.17\\0.13\\0.01\\0.10\\0.06\\0.13\\0.01\\0.09\end{pmatrix}+.5\times\begin{pmatrix}0.41\\0.33\\0.26\\0\\0\\0\\0\\0\\0\end{pmatrix}\tag{4.7}$$

对于给定的指称项 m，基于图的协同推断方法用公式4.8计算 m 对应的预测目标实体。其中 $r(e)$ 是随机行走达到稳定状态后，候选实体 e 通过图推断后得到的评分，与局部相似度 $CP(m,e)$ 相乘后得到综合评分，并取指称项 m 对应的候选实体集中，综合评分最高的候选实体作为 m 的预测目标实体。

$$m.e^*=\underset{e}{\text{argmax}}\,CP(m,e)\times r(e)\tag{4.8}$$

表4.2给出了本节的例子中，各个候选实体在随机行走达到稳定状态后的推断评分。通过观察可以发现，经过协同推断进程后，候选实体 *Michael Jordan*、*Chicago Bulls*、*Space Jam* 是在其对应的指称项的候选实体集里面推断评分最高的，通过公式4.8计算得到的综合评分也是最高的，验证了基于图的协同推断方法的有效性。

表 4.2: 主动学习策略结果评价

候选实体	**Michael I. Jordan**	**Michael Jordan**	**Michael B. Jordan**
$r(e)$	0.01	0.10	0.06
候选实体	**Chicago Bulls**	**Bulls**	**Space Jam**
$r(e)$	0.13	0.01	0.09

4.2 银标准语料构建方法

银标准语料即存在错误标注，但不会对模型训练产生很大影响的语料。在成本有限的情况下，提高银标准语料质量对模型训练有重要意义。虽然 Han 等人[18] 通过实验证明了基于图的协同推断方法是目前性能较好的实体链接方法，但是该方法在本文所使用的数据集上仅能达到 73.3% 的正确率。其在正确率上距离高质量银标准语料还有一定差距。常用的方法是人工对部分未标注样本进行标注，然后通过已标注样本对未标注样本进行证据传播，从而提高其它未标注样本的预测精度。

在人工标注并证据传播的过程中，有以下两个关键问题：

（1）如何从未标注指称项样本中选择需要交由人工标注的待标注样本。

（2）对指称项对应的目标实体进行标注以后，如何影响其它未标注样本的目标实体预测。

本文将在以下两节分别介绍待标注指称项的选择方法和已标注指称项的证据传播方法。

4.2.1 待标注指称项选择方法

以下是四种待标注指称项选择方法，其中第一种采用顺序选择方法，后三种都是基于主动学习的选择方法。

4.2.1.1 顺序指称项选择法

该方法是一种基线方法。对一篇文档 D 中的所有指称项 $M = \{m_1, m_2, ..., m_n\}$，按照指称项出现的先后顺序对指称项进行标注。

该方法的优点在于实现简单，并且顺序标注符合人的标注习惯。该方法的缺点是，当实体链接模型性能较高的时候，提交人工标注的指称项有很大概率已经被正确预测，增加人工标注的工作量，标注效率较低。

4.2.1.2 最大不确定度的指称项选择法

该方法利用基于间隔的主动学习方法对待标注指称项进行了选择。对于给定指称项 m 及其候选实体集合 E_m，对 E_m 中的每个候选实体 e 的综合评分做归一化处理，得到

候选实体 e 是目标实体的置信度，如公式4.9所示。

$$P(e|m) = \frac{CP(m,e) \times r(e)}{\sum_{e \in E_m} CP(m,e) \times r(e)} \tag{4.9}$$

通过公式4.10计算指称项 m 的候选实体中综合评分最高的两个候选实体的置信度的差的绝对值，以此作为指称项 m 被正确链接的置信度。

$$Confidence(m) = P(e^{*1}|m) - P(e^{*2}|m)$$

其中，

$$\begin{aligned} e^{*1} &= \underset{e_i \in E_m}{\text{argmax}}\, P(e_i|m) \\ e^{*2} &= \underset{e_i \in E_m \setminus e^{*1}}{\text{argmax}}\, P(e_i|m) \end{aligned} \tag{4.10}$$

得到指称项 m 被正确链接的置信度以后，通过公式4.11计算指称项 m 链接的不确定度。

$$Uncertainty(m) = 1 - Confidence(m) \tag{4.11}$$

链接不确定度越大的指称项，预测标注错误的可能性越大。因此，在每一轮待标注样本选择过程中，需要选择链接不确定度最大的指称项交由人工标注，通过这种方式尽可能找出错误预测的指称项，从而提高人工标注的效率。

4.2.1.3 基于同名指称项的最大标注回报选择法

该方法在利用主动学习方法选择待标注样本的过程中，不仅考虑到了样本的不确定度，同时也考虑到了样本的代表性。在同一篇文档中，相同的指称项可能出现多次，这些指称项很可能指向同一个实体。对这些指称项中的某一个指称项做人工标注，可以提高其它指称项预测的正确率。这里，对指称项 m 进行标注的回报率进行评估时，评估结果由与 m 同名的指称项集合 $M_m = \{m' | Name_m = Name_{m'}\}$ 中的所有指称项共同确定，如公式4.12所示。

$$Reward(M_m) = \sum_{m' \in M_m} Uncertainty(m') \tag{4.12}$$

从公式中可以直观地看出，同一文档中，同名指称项出现的次数越多，各个指称项链接的不确定度越大，则对指称项进行人工标注的回报率越大。对 $Reward(M_m)$ 值最大的指称项集合 M_m 中的某个指称项进行人工标注，理论上对语料库实体链接预测正确率的提升效果最大。

4.2.1.4　基于相似指称项的最大标注回报选择法

存在一些指称项集合，它们存在共指关系，但是它们名字的字符串并不是严格相等的，可能是缩写、别名等其它形式。例如在同一篇文档中，指称项“Microsoft”和指称项“MS”可能就是全名和缩写的关系，它们所指向的实体是相同的。基于字符串完全匹配的方法的缺点是无法利用这类指称项之间的共指关系。

为了克服上述缺点，本文提出了基于相似指称项的最大标注回报选择方法。本文通过知识库相关信息抽取得到候选实体词典，然后借助候选实体词典，根据指称项对应的候选实体集计算不同指称项之间的语义相似度。

$$MS(m_1, m_2) = \frac{|E_{m_1} \cap E_{m_2}|}{\min(|E_{m_1}|, |E_{m_2}|) + 0.1} \tag{4.13}$$

公式4.13是指称项 m_1 和指称项 m_2 之间语义相似度的计算方法。该公式的假设是，相同或相似的指称项，对应的候选实体集相似度也应该会更高。在极端情况下，两个相同的指称项，候选实体集完全相同，指称项之间的语义相似度为 1。相反，完全不相关的指称项，对应的候选实体集交集为空，则指称项之间的语义相似度为 0。

因此，在评估指称项 m 的标注回报率时，评估结果由与 m 相似的指称项集合 $M_m = \{m^{'}|MS(m, m^{'}) >= \alpha\}$（公式中的 α 通过实验获得，本文取值为 0.8）中的所有指称项共同确定，计算方式与公式4.12相同，也是由指称项集合 M_m 中的所有指称项链接不确定度的和得到。

同样地，选择需要人工标注的指称项时，也是需要从 $Reward(M_m)$ 最大的指称项集合中取出一个指称项进行标注，这里从集合中选取的原则是选择不确定度最大的指称项。

4.2.2　已标注指称项证据传播方法

同一篇文档会包含多个指称项，一方面，这些指称项通常会有相互关联的关系，例如同一篇文档中既包含“Jordan”这个指称项，又包含“Bulls”这个指称项。如果人工标注“Jordan”指向 *Michael Jordan* 这个实体，那么“Bulls”则很可能指向 *Chicago Bulls* 这个实体。另一方面，采用基于同名指称项的最大标注回报选择法和基于相似指称项的最大标注回报选择法时，经过人工标注的指称项会对文档中的其它未标注指称项的目标实体预测产生一定的影响。

在同一篇文档中，相同或者相似的指称项可能会出现多次，例如出现多个“Jordan”指称项，当人工标注其中一个“Jordan”指向 *Michael Jordan* 这个实体，那么其它“Jordan”指称项很可能也指向 *Michael Jordan* 这个实体。另外还需要考虑到不同字符串表示的指称项可能存在共指关系，即它们可能指向同一个实体，例如指称项“Jordan”和指称项“Michael Jordan”如果出现在同一篇文档中，则它们也很可能指向同一个实体 *Michael Jordan*。

4.2.2.1 图的迭代推断

当标注者标注指称项 m 指向实体 e 以后，需要对协同推断图做以下处理。首先，需要删除指称项 m 对应的候选实体集 E_m 中除目标实体 e 以外的其他候选实体节点。然后删除与被移除节点相连的边。

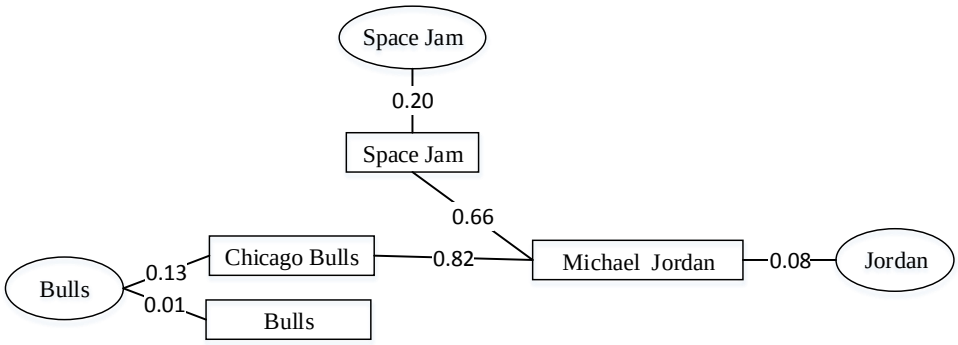


图 4.3: 协同推断图示例图

例如将指称项“Jordan”进行人工标注指向 *Michael Jordan* 实体，改动后的协同推断图如图4.3所示。图中指称项“Jordan”对应的候选实体集中 *Michael I. Jordan* 实体和 *Michael B. Jordan* 实体对应的节点被删除，指称项“Jordan”连接到实体 *Michael I. Jordan* 和实体 *Michael B. Jordan* 的边被删除，另外，实体 *Michael I. Jordan* 和实体 *Michael B. Jordan* 节点连接到其它实体节点的边也被删除。

由于经过人工标注以后协同推断图发生了变化，因此对应到协同推断过程，需要根据改动以后的协同推断图对初始证据向量 s 和证据传递矩阵 T 进行改动。改动后的初始证据向量 s 和证据传递矩阵 T 如分别如公式4.14和公式4.15所示。

$$s = (0.41, 0.33, 0.26, 0, 0, 0, 0)^T \tag{4.14}$$

$$T = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0.92 & 0 & 0.55 & 0 & 0 & 0 \\ 0 & 0.08 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0.45 & 0 & 0 & 0 \end{pmatrix} \tag{4.15}$$

通过观察可以发现，由于删除了部分节点，s 和 T 的维度发生了变化，另外协同推断图边的权值变化也引起了 T 中部分值的变化。后续推断过程按照第4.1.4节描述的方法进行即可。

4.2.2.2　标注证据传播

当人工标注指称项 m 指向目标实体 e 以后，需要将标注结果传播到同一篇文档中的其他可能共指的指称项。本文采用了两种标注证据传播方法：基于字符串完全匹配的标注证据传播方法和基于相似指称项的标注证据传播方法。

（1）基于字符串完全匹配的标注证据传播

当人工标注文档 D 中的某一个指称项 m 指向目标实体 e 以后，在文档 D 中搜索与指称项 m 名字完全相同的未标注指称项集合 $M_m = \{m^{'}|Name_m = Name_{m^{'}}\}$。然后需要调整证据传递矩阵 T 中所有 $m^{'} \in M_m$ 的指称项节点连接到其对应候选实体节点的边的权值，如公式4.16所示。

$$T_{ij} = \begin{cases} 1 & \text{节点}j\text{对应指称项}m^{'}\text{，节点}i\text{对应实体}e \\ 0 & \text{节点}j\text{对应指称项}m^{'}\text{，节点}i\text{对应除}e\text{以外的其它实体} \end{cases} \tag{4.16}$$

（2）基于相似指称项的标注证据传播

当人工标注文档 D 中的某一个指称项 m 指向目标实体 e 以后，在文档 D 中搜索与指称项 m 语义相似度不小于 α 的未标注指称项集合 $M_m = \{m^{'}|MS(m, m^{'}) >= \alpha\}$。同时依然要对证据传递矩阵 T 做调整，按照公式4.17对所有 $m^{'} \in M_m$ 的指称项节点连接到实体节点的边的权值做调整。

$$T_{ij} = \begin{cases} MS(m, m^{'}) \times T_{ij} & \text{节点}j\text{对应指称项}m^{'}\text{，} \\ & \text{节点}i\text{对应实体}e \\ (1 - MS(m, m^{'})) \times T_{ij} & \text{节点}j\text{对应指称项}m^{'}\text{，} \\ & \text{节点}i\text{对应除}e\text{以外的其它实体} \end{cases} \tag{4.17}$$

调整权值之后，还需要对所有指向指称项的节点 j 通过公式4.18对权值做归一化处理。

$$T_{ij} = \frac{T_{ij}}{\sum_k T_{kj}} \tag{4.18}$$

4.3　实验结果与分析

4.3.1　实验环境

同3.5.1节。

4.3.2　实验数据集

数据集同3.5.2节，因为本章实验内容为实体链接银标准语料的辅助标注，因此不需要将数据集划分为训练集和测试集。

4.3.3 实验设置

本章研究主动学习方法在实体链接银标准语料构建中的作用，并验证本文提出的融合证据传播的主动学习方法对减少人工标注样本工作量的效果。实验过程中不直接使用 Aida 数据集的标注结果，而只使用模型选择的待标注样本的标注结果。本章设计了以下六个对比实验:

（1）以顺序标注方法作为基线方法，评价标注过程中，语料中指称项链接的正确率的变化，同下列基于主动学习的标注方法做比较。

（2）以最大不确定度的指称项选择法选择待标注指称项，在标注证据传播过程中，不对当前标注结果进行传播。

（3)）以最大不确定度的指称项选择法选择待标注指称项，在标注证据传播过程中，基于字符串完全匹配对当前标注结果进行传播。

（4）以最大不确定度的指称项选择法选择待标注指称项，在标注证据传播过程中，基于相似指称项对当前标注结果进行传播。

（5）以基于同名指称项的最大标注回报选择法选择待标注指称项，在标注证据传播过程中，基于字符串完全匹配对当前标注结果进行传播。

（6）以基于相似指称项的最大标注回报选择法选择待标注指称项，在标注证据传播过程中，基于相似指称项对当前标注结果进行传播。

4.3.4 实验结果与分析

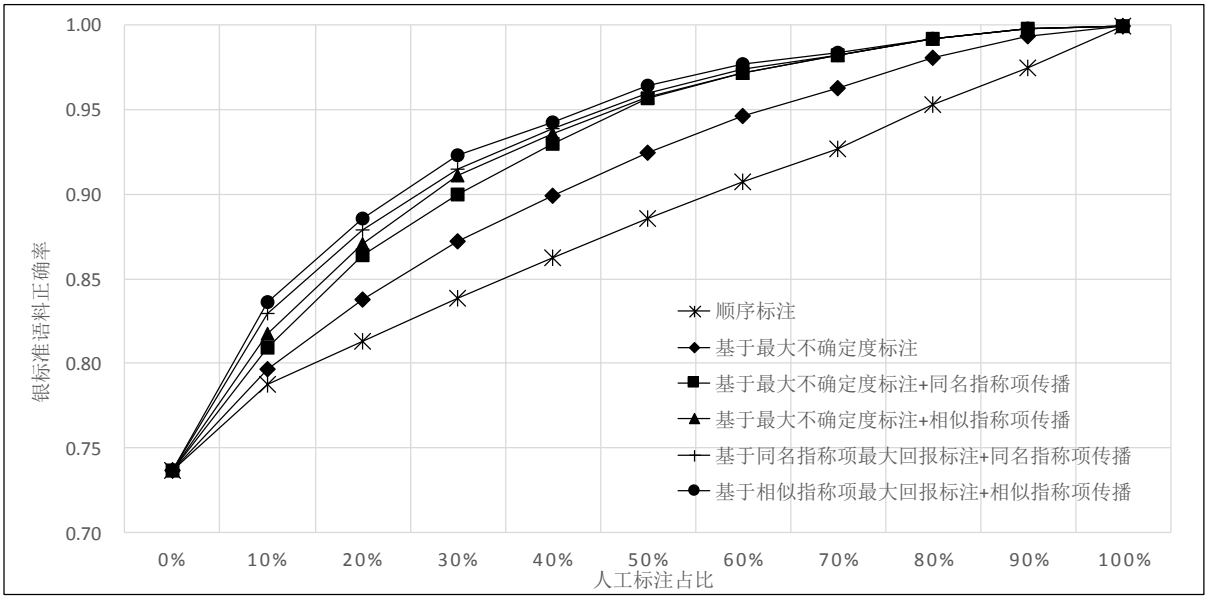


图 4.4: 银标准语料库构建实验结果

本文实验环节测试了上述六种标注方式在数据集上的标注效果。如图4.4所示。顺序标注并且没有对标注证据进行传播的方法，标注效率是最低的。以最大不确定度的指

称项选择法选择待标注指称项，在不进行标注证据传播的情况下，标注效率明显优于顺序标注的方式。这是因为该基于主动学习的方法能够有效搜索出预标注不确定度大的样本，保证交由人工标注的样本很可能是预标注错误的样本，从而提高标注效率，并且这一假设通过实验得到了证明。

以最大不确定度的指称项选择法选择待标注指称项，并基于同名指称项和基于相似指称项的方式对当前标注结果进行标注证据传播，观察曲线可以发现，相比于不对标注证据传播的方式，标注效率得到了显著提升，并且基于相似指称项的方式优于基于字符串完全匹配的方式。这是因为，同一文档中相同指称存在多次出现的情况，基于同名指称项的标注证据传播能利用一次标注提高多个未标注指称项的链接正确率。另外，基于相似指称项的证据传播方式能将标注结果传播到存在共指关系的未标注指称项，因此比基于同名指称项的标注证据传播方法性能更好。实验证明，在实体链接银标准语料标注任务中，融合标注证据传播，对提升主动学习方法的性能有显著的效果。

在待标注样本选择阶段，基于同名指称的最大标注回报选择法和基于相似指称的最大标注回报选择法，相比最大不确定度的指称项选择法，标注效率得到提升。提升的原因是，在样本选择的阶段，不仅考虑到单个指称项的预测链接不确定度，还考虑了标注该指称项对文档中其它存在共指关系的指称的影响。

综合分析实验结果数据，在仅标注 50% 的指称项时，顺序标注的方式只能将语料库标注正确率提升到 88.6%，而以基于相似指称的最大标注回报选择法选择待标注指称项，在标注证据传播过程中，基于相似指称项对当前标注结果进行传播的方法，标注正确率提升到了 96.4%。

4.4 本章小结

本章介绍了一种基于主动学习的实体链接银标准语料库构建方法。并且，本文基于实体链接任务的特点，对已有的主动学习方法进行了改进，包括引入了基于标注回报率待标注样本选择的评价方式，以及加入了标注样本的证据传播方法，提升了主动学习方法的性能。在本章实验环节，本文对提出的这些方法进行了验证，并分析了实验结果。

第五章 总结与展望

5.1 总结

本文研究基于主动学习的实体链接方法，解决当前存在的两个问题，并通过实验进行验证。

（1）使用监督学习模型处理实体链接任务时，需要人工标注足够的训练样本集，此项工作费时费力。

基于主动学习的实体链接监督模型训练方法可有效解决此问题。使用基于流行度的初始样本集选择方法提高初始分类器的性能，使用综合不确定度和流行度的样本选择方法加快迭代训练过程中模型的收敛速度。

在基于主动学习的实体链接监督模型训练方法的实验环节中，给定初始样本集较小时，本文提出的初始样本选择方法的初始分类器性能比随机选择的基线方法的性能提升了 10.1%。在后续迭代训练过程中，本文提出的综合不确定度和流行度的样本选择方法相比传统的基于不确定度的选择算法的 $deficiency$ 值降低了 16.1%。

（2）在成本有限的情况下，传统的标注方法对于构建高质量的实体链接银标准语料，效率低下。

可采用基于主动学习的实体链接银标准语料构建来解决此问题。借助基于图的协同推断方法对未标注样本中的指称项进行预标注，然后通过主动学习方法选择预标注不确定度高的指称项进行人工标注，以此提高错误预标注的命中率，提高人工标注效率。另一方面，通过标注证据传播，将人工标注结果传播到未标注指称项，从而减少人工标注的工作量。

在基于主动学习的实体链接银标准语料构建方法的实验环节中，本文对主动学习的待标注样本选择方法进行了改进，并融入了标注证据传播，加速了主动学习进程。在仅标注 50% 的指称项的情况下，实验中性能最佳的方法相比于非主动学习的顺序标注方式，标注语料正确率提升了约 8 个百分点。

5.2 未来展望

从实验结果看，本文提出了一系列基于主动学习的实体链接方法，但仍存在值得改进的地方。

（1）在基于有监督学习的实体链接模型训练过程中，考虑到使用的特征比较简单，后续工作可以考虑用分布式词表示方法计算指称项与候选实体之间的相似度。

（2）实体链接模型采用的是单个模型的训练方式，模型性能在一定程度上受到模型选择的制约。未来工作中，我们将会使用多个模型来处理实体链接任务，并基于委员会

的主动学习方法加速模型的训练，利用各种模型的差异来解决单个模型具有制约性的缺陷。

（3）在基于主动学习的实体链接银标准语料构建方法中，本文主要考虑的是文档内指称项之间的关系，文档与文档之间是相互独立的。而实际上，不同文档之间的指称项也可能存在相互关联的关系。如何处理这种跨文档的指称项关系检测，对提升语料标注效率具有重要意义，这也是未来工作之一。

致谢

时光匆匆而逝，我的硕士三年研究生生活即将结束。回首这三年，心中倍感充实。我要感谢我的母校，我庆幸我是一名东大的学子，学校浓厚的学术氛围和舒适的学习环境将令我终生难忘。

首先我要感谢我的导师高志强教授。他在忙碌的教学和科研工作中挤出时间来审查我的开题报告，修改我的论文，是高老师的意见和指导帮助我顺利完成了研究课题。此外，从我进入实验室起，高老师严谨细致、一丝不苟的学术作风也深深地影响着我。同时，也感谢实验室的所有同门朱曼、刘倩、鲁廷明、全志斌、归耀城、李雪莲、吕永涛、王辰、倪朝曦、潘敬敏、司马强、王煜、王李荣、余云秀、张赏、刘金晶、范云龙、李斌、刘延栋、汪文涛、衣克买提等，尤其是鲁廷明师兄、全志斌师兄、余云秀师妹，他们在我的研究和论文上提出了诸多宝贵的意见，并在精神上给予我诸多鼓励，使我受益良多。

感谢我在东大相识的所有伙伴温潇、罗鸿飞等，你们为我的校园生活带来了欢乐与感动。感谢我的室友们在日常生活中对我的关心。感谢所有在学习与项目中与我共同奋斗过的朋友，你们不仅让我体会到了成功的喜悦，也收获了真挚的友情。也许在不久的将来我们会天各一方，但是我会永远记住你们，我的伙伴们。

最后，我要感谢我的父母，让我在人生漫漫长路中有了厚实的依靠。养育之恩，无以回报，在未来的日子里，我会更加努力学习与工作，不辜负父母对我的殷切期望。

吴展鹏

2017 年 5 月于南京

参考文献

[1] Ghosh, Saptarshi and Ghosh, Kripabandhu. Overview of the FIRE 2016 Microblog track: Information Extraction from Microblogs Posted during Disasters[C]. 2016. 7–10.

[2] Dredze, Mark, McNamee, Paul, Rao, Delip, et al. Entity disambiguation for knowledge base population[C]. In: Proceedings of the 23rd International Conference on Computational Linguistics. 2010. 277–285.

[3] Yih, Scott Wen-tau, Chang, Ming-Wei, He, Xiaodong, et al. Semantic parsing via staged query graph generation: Question answering with knowledge base[EB/OL]. https://www.microsoft.com/en-us/research/wp-content/uploads/2016/02/ACL15-STAGG.pdf. 2015.

[4] Liu, Xiaohua, Li, Yitong, Wu, Haocheng, et al. Entity Linking for Tweets.[C]. In: ACL (1). 2013. 1304–1311.

[5] Zheng, Zhicheng, Li, Fangtao, Huang, Minlie, et al. Learning to link entities with knowledge base[C]. In: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. 2010. 483–491.

[6] Fu, JinLan, Qiu, Jie, Guo, Yunlong, et al. Entity linking and name disambiguation using SVM in chinese micro-blogs[C]. In: Natural Computation (ICNC), 2015 11th International Conference on. 2015. 468–472.

[7] Kim, Youngsik and Choi, Key-Sun. Entity Linking Korean Text: An Unsupervised Learning Approach using Semantic Relations.[C]. In: CoNLL. 2015. 132–141.

[8] Hakkani-Tür, Dilek, Heck, Larry, and Tur, Gokhan. Using a knowledge graph and query click logs for unsupervised learning of relation detection[C]. In: Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on. 2013. 8327–8331.

[9] Chen, Yukun, Lasko, Thomas A, Mei, Qiaozhu, et al. A study of active learning methods for named entity recognition in clinical text[J]. Journal of biomedical informatics, 2015, 58:11–18.

[10] Cormack, Gordon V and Grossman, Maura R. Scalability of Continuous Active Learning for Reliable High-Recall Text Classification[C]. In: Proceedings of the 25th ACM International on Conference on Information and Knowledge Management. 2016. 1039–1048.

[11] Chen, Yukun, Cao, Hongxin, Mei, Qiaozhu, et al. Applying active learning to supervised word sense disambiguation in MEDLINE[J]. Journal of the American Medical Informatics Association, 2013, 20(5):1001–1006.

[12] Hu, Weiming, Hu, Wei, Xie, Nianhua, et al. Unsupervised active learning based on hierarchical graph-theoretic clustering[J]. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 2009, 39(5):1147–1161.

[13] Zhang, Tao, Liu, Kang, Zhao, Jun, et al. Cross Lingual Entity Linking with Bilingual Topic Model.[C]. In: IJCAI. 2013.

[14] Strauss, Benjamin, Toma, Bethany E, Ritter, Alan, et al. Results of the wnut16 named entity recognition shared task[C]. In: Proceedings of the 2nd Workshop on Noisy User-generated Text. 2016. 138–144.

[15] Quimbaya, Alexandra Pomares, Múnera, Alejandro Sierra, Rivera, Rafael Andrés González, et al. Named Entity Recognition Over Electronic Health Records Through a Combined Dictionary-based Approach[J]. Procedia Computer Science, 2016, 100:55–61.

[16] Lu, Tingming, Zhu, Man, and Gao, Zhiqiang. Reducing Human Effort in Named Entity Corpus Construction Based on Ensemble Learning and Annotation Categorization[C]. In: International Conference on Computer Processing of Oriental Languages. 2016. 263–274.

[17] Hachey, Ben, Radford, Will, Nothman, Joel, et al. Evaluating entity linking with Wikipedia[J]. Artificial intelligence, 2013, 194:130–150.

[18] Han, Xianpei, Sun, Le, and Zhao, Jun. Collective entity linking in web text: a graph-based method[C]. In: Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval. 2011. 765–774.

[19] Rao, Delip, McNamee, Paul, and Dredze, Mark. Entity linking: Finding extracted entities in a knowledge base[EB/OL]. 2013.

[20] Shen, Wei, Wang, Jianyong, and Han, Jiawei. Entity linking with a knowledge base: Issues, techniques, and solutions[J]. IEEE Transactions on Knowledge and Data Engineering, 2015, 27(2):443–460.

[21] Liu, Xiaohua, Li, Yitong, Wu, Haocheng, et al. Entity Linking for Tweets.[C]. In: ACL (1). 2013. 1304–1311.

[22] Fu, JinLan, Qiu, Jie, Guo, Yunlong, et al. Entity linking and name disambiguation using SVM in chinese micro-blogs[C]. In: Natural Computation (ICNC), 2015 11th International Conference on. 2015. 468–472.

[23] Zhao, YaHui, Li, Haodi, Chen, Qingcai, et al. ICRC-DSEDL: A Film Named Entity Discovery and Linking System Based on Knowledge Bases[EB/OL]. 2016.

[24] Cossock, David and Zhang, Tong. Subset ranking using regression[C]. In: International Conference on Computational Learning Theory. 2006. 605–619.

[25] Yuan, Ke, Gao, Liangcai, Wang, Yuehan, et al. A mathematical information retrieval system based on RankBoost[C]. In: Digital Libraries (JCDL), 2016 IEEE/ACM Joint Conference on. 2016. 259–260.

[26] Chen, Jing, Xiong, Chenyan, and Callan, Jamie. An Empirical Study of Learning to Rank for Entity Search[C]. In: Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval. 2016. 737–740.

[27] Song, Yang, Wang, Hongning, and He, Xiaodong. Adapting deep ranknet for personalized search[C]. In: Proceedings of the 7th ACM international conference on Web search and data mining. 2014. 83–92.

[28] Xu, Jun and Li, Hang. Adarank: a boosting algorithm for information retrieval[C]. In: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval. 2007. 391–398.

[29] Yue, Yisong, Finley, Thomas, Radlinski, Filip, et al. A support vector method for optimizing average precision[C]. In: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval. 2007. 271–278.

[30] Cao, Zhe, Qin, Tao, Liu, Tie-Yan, et al. Learning to rank: from pairwise approach to listwise approach[C]. In: Proceedings of the 24th international conference on Machine learning. 2007. 129–136.

[31] Burges, Christopher JC. From ranknet to lambdarank to lambdamart: An overview[J]. Learning, 2010, 11(23-581):81.

[32] Hanghang, Tong, Christos, Faloutsos., and Jia-yu, Pan. Fast random walk with restart and its applications[J]. 2006.

[33] Hoffart, Johannes. Discovering and disambiguating named entities in text[C]. In: Proceedings of the 2013 SIGMOD/PODS Ph. D. symposium. 2013. 43–48.

[34] Tjong Kim Sang, Erik F and De Meulder, Fien. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition[C]. In: Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4. 2003. 142–147.

[35] Ji, Heng, Grishman, Ralph, Dang, Hoa Trang, et al. Overview of the TAC 2010 knowledge base population track[C]. In: Third Text Analysis Conference (TAC 2010). 2010. 3–3.

[36] Lewis, David D and Catlett, Jason. Heterogeneous uncertainty sampling for supervised learning[C]. In: Proceedings of the eleventh international conference on machine learning. 1994. 148–156.

[37] Muslea, Ion, Minton, Steven, and Knoblock, Craig A. Active learning with multiple views[J]. Journal of Artificial Intelligence Research, 2006, 27:203–233.

[38] Freund, Yoav, Seung, H Sebastian, Shamir, Eli, et al. Selective sampling using the query by committee algorithm[J]. Machine learning, 1997, 28(2):133–168.

[39] Chen, Yukun, Lasko, Thomas A, Mei, Qiaozhu, et al. A study of active learning methods for named entity recognition in clinical text[J]. Journal of biomedical informatics, 2015, 58:11–18.

[40] Cormack, Gordon V and Grossman, Maura R. Scalability of Continuous Active Learning for Reliable High-Recall Text Classification[C]. In: Proceedings of the 25th ACM International on Conference on Information and Knowledge Management. 2016. 1039–1048.

[41] Alonso, Héctor Martínez, Plank, Barbara, Johannsen, Anders, et al. Active learning for sense annotation[C]. In: Proceedings of the 20th Nordic Conference of Computational Linguistics, NODALIDA 2015, May 11-13, 2015, Vilnius, Lithuania. 2015. 245–249.

[42] Ayache, Stéphane and Quénot, Georges. Video corpus annotation using active learning[C]. In: European Conference on Information Retrieval. 2008. 187–198.

[43] 李航. 统计学习方法 [M]. 清华大学出版社, 北京, 2012. 95–114.

[44] Ravuri, Suman and Stolcke, Andreas. A comparative study of recurrent neural network models for lexical domain classification[C]. In: Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on. 2016. 6075–6079.

[45] Settles, Burr. Active learning[J]. Synthesis Lectures on Artificial Intelligence and Machine Learning, 2012, 6(1):1–114.

[46] Li, Huiying and Shi, Jing. Linking Named Entity in a Question with DBpedia Knowledge Base[C]. In: Joint International Semantic Technology Conference. 2016. 263–270.

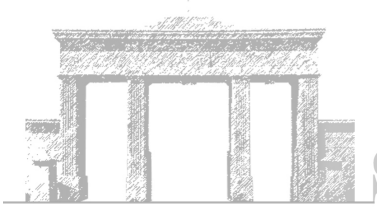
[47] Schein, Andrew I and Ungar, Lyle H. Active learning for logistic regression: an evaluation[J]. Machine Learning, 2007, 68(3):235–265.

[48] 张涛, 刘康, and 赵军. 一种基于图模型的维基概念相似度计算方法及其在实体链接系统中的应用 [J]. 中文信息学报, 2015, 29(2):58–67.

[49] Pedersen, Ted, Purandare, Amruta, and Kulkarni, Anagha. Name discrimination by clustering similar contexts[C]. In: International Conference on Intelligent Text Processing and Computational Linguistics. 2005. 226–237.

[50] Hu, Jian, Wang, Gang, Lochovsky, Fred, et al. Understanding user's query intent with wikipedia[C]. In: Proceedings of the 18th international conference on World wide web. 2009. 471–480.

心求至善



SOUTHEAST UNIVERSITY