毕业设计（论文）报告

题目　基于多知识库的表格实体链接系统

计算机科学与工程 院（系）计算机科学与技术 专　业

学　　号　09013119

学生姓名　严晟嘉

指导教师　李慧颖

顾问老师　漆桂林

起讫日期　2017.02.20 —2017.05.22

设计地点　东南大学九龙湖校区

摘　要

在当今的万维网上包含着超大规模的蕴含丰富价值的关系型数据，而这些关系型数据大多以 HTML 表格 (也就是 Web 表格) 的形式呈现。抽取 Web 表格中的语义来制造机器可以理解的知识如今已经成为了一个热门的研究领域。从 Web 表格中抽取出丰富且高质量的语义信息是一个充满挑战和价值的任务，其中关键一步就是实体链接 (Entity Linking)，其目的在于建立表格中的字符串指称 (Mention) 到知识库中的相应的参考实体 (Entity) 的链接。

目前大部分的实体链接都是使用单个特定的跨领域知识库来作为实体数据的来源。但由于单个知识库中的实体数量有限且对实体的覆盖范围不广，从而导致单知识库实体链接中实体缺失和链接错误的问题常有发生。在本文中，我会介绍我的本科毕业设计“基于多知识库的表格实体链接系统” 的设计思路和具体实现。这个系统中包含了 2 种通用的表格实体链接的新方法。与先前的工作不同的是，这 2 个方法都使用多知识库而不是单知识库作为实体数据的来源，来提升表格实体链接的质量。

方法一是一个 “两步走” 的方法，首先将一个基于概率图模型的随机游走算法应用于单知识库的表格实体链接，然后使用多知识库间已有的和新学习到的 “sameAs” 关系来提升前一步中实体链接结果的质量。方法二融合了方法一中的两步，修改了方法一中概率图模型中的节点定义，直接将多知识库间已有的和新学习到的 “sameAs” 关系融合进了概率图模型，从而实现了一步到位的效果。

我设计了一些对比实验来比较本文提出的 2 个方法与目前最先进的实体链接系统 (TabEL, LIEGE) 的效果差异。所有实验都是在基于 Zhishi.me 人工标注的 Web 表格上进行的，Zhishi.me 涵盖了三个最大的中文百科知识库：百度百科，互动百科和中文维基百科。实验表明，文中提出的 2 个方法在准确率、召回率、F1 值等评价指标上表现得都很不错。除此之外，系统在实体消岐时引入了更多的有效的特征，还增加了一些有意义的功能，比如实体链接和 “sameAs” 关系的迭代学习。实验中用到的已人工标注实体的表格语料库和整个系统的代码都是开放的，供未来的研究工作使用。

关键词： 实体链接，Web 表格，多知识库

Abstract

The World-Wide Web contains a huge scale of valuable relational data, which are embedded in HTML tables (i.e. Web tables). It's a challenging and valuable task that extracting abundant and high-quality semantics of Web tables is. A key step in extracting the semantics of Web tables is Entity Linking (EL), which aims to map the string mentions in table cells to their referent entities in a Knowledge Base (KB).

At present most entity linking researches use single cross-domain KBs as source of entities. However, the quantity of entities and the coverage in a single KB is limited. Thus, the issues of entity absence and linking error often appear in the process of EL in a single KB. In this paper, I will introduce the design ideas and concrete implementations of my undergraduate graduate project "Entity Linking System in Web Tables with Multiple Linked Knowledge Bases", which includes two new approaches for EL in Web tables. Different from previous work, the proposed approaches replace a single KB with multiple linked KBs as the sources of entities to improve the quality of EL.

Approach One is a two-step method. I first apply a random walk algorithm based on a general probabilistic graphical model to EL in Web tables with each single KB. Then, I leverage the existing and newly learned "sameAs" relations between the entities from different KBs to help improve the results of EL in the first step. Approach Two merges two steps from Approach One. I modify node definitions in the probabilistic graphical model in Approach One and directly add the "sameAs" relations between the entities from different KBs into the model.

Some comparison experiments are designed to compare the effect of my system proposed in this paper with the state-of-the-art table's EL systems (TabEL, LIEGE). All experiments are conducted on the sampled Web tables with Zhishi.me, which consists of three biggest linked Chinese encyclopedic KBs: Baidubaike, Hudongbaike and Zhwiki. The experimental results show that my approaches have a good performence in different evaluation metrics, such as precision, recall and F1 value. Besides, the system adds some meaningful functions, such as iterative learning between EL and "sameAs" relations. I also make my manually annotated table corpus and the code of the entire system publicly available for future work.

KEY WORDS: Entity Linking, Web Tables, Multiple Linked Knowledge Bases

目 录

第一章　绪论

1.1 研究动机

如今 Web 上的内容每天都在以指数级的速度增长[1]，这也使得 Web 近年来成为世界上最大的数据集散地之一。据估计 Web 上有超过 141 亿张表格，其中 1.54 亿张表格包含关系型数据并且仅 Wikipedia 就是大约 160 万关系型表格的来源。可见 Web 表格，换句话说是 Web 上的 HTML 表格，是关系型数据的一个重要来源和信息抽取 (Information Extraction) 系统的一个重要输入。与普通文本不同，单张关系型表格就包含一系列高质量的关系实体并且在表格的列头中包含与实体相关的元数据。Web 上包含关系型数据的表格中蕴含的巨大财富和价值使得表格的语义解释 (Semantic Interpretation)，也就是将 Web 表格转换成机器能够理解的知识这一任务成为热门的研究领域。

另一方面，诸如维基百科等知识共享社区的蓬勃发展和信息抽取技术的进步已经促成了大规模机器可读知识库的自动化建设。目前世界上已经出现了上百个领域不同、规模不一的知识库并且它们的规模每天都在飞速增长。知识库中包含着整个世界的实体，实体的语义类别及其相互关系的丰富信息。这样典型的例子包括 YAGO[2]，DBpedia[3] 和 Freebase[4]。在中文知识库中较有影响力的就是由上海交通大学和东南大学共同建立的 zhishi.me[5]。

Web 上虽然蕴含着许多有价值的数据，但更多的却是各种原始并且充斥着噪声的数据，其中有些甚至是错误的。这些数据大都是以自然语言的形式存在，然而由于自然语言表达的多样性与歧义性，使得它们很难被计算机直接处理或者理解。面对如此规模庞大而嘈杂的数据，信息过载的现象每天都在发生。信息过载意味着在找到想要的有用的信息之前，需要处理大量的无用数据，计算机获取有效信息的效率常常受到限制。为了减轻信息过载带来的负面影响以及处理自然语言的多样性问题和歧义性问题，语义网 (Semantic Web) 的概念应运而生，旨在对现有万维网上的文档进行元数据 (Meta Data) 标注，使计算机能够理解词语和概念以及它们之间的逻辑关系。将 Web 数据与知识库链接起来是非常有利于标注 Web 上的大规模数据的，并且有助于实现语义网的愿景[6]。许多为了解释说明 Web 表格内含的语义的研究工作[1][7][8][9][10][11] 是将其内容标注成 RDF 三元组。这种语义标注 (Semantic Annotation) 的关键一步就是实体链接 (Entity Linking)，将出现在 Web 表格单元格中的字符串指称 (Mention) 链接到其在给定知识库中对应的参考实体 (Entity)。

实体链接技术的发展可以带动许多不同的应用的发展，比如知识库补全，自然语言问答系统和语义搜索系统。随着社会的发展，新的知识被创造出来并以数据的形式表现

在 Web 上。因此，利用这些新知识扩充现有的知识库显得越发重要。然而，为了将这些新知识插入到现有知识库中，会不可避免地需要一个系统，来将字符串指称，也就是与已抽取出的知识相关联的指称，链接到知识库中的相应实体。例如，自然语言问答系统依靠它们支持的知识库来回答用户的问题。为了回答“苹果公司创始人史蒂夫·乔布斯的诞生日期”的问题，该系统应首先利用实体链接技术将查询语句中的“史蒂夫·乔布斯”映射到美国企业家，而不是美国传记电影，然后从知识库中直接取回名为“史蒂夫·乔布斯”的出生日期。除此之外，实体链接对数据集成很有帮助，可以将不同页面、文档和站点上的实体信息进行集成。可见，Web 表格上的实体链接技术很有价值并且拥有广阔的应用前景。

1.2 研究现状

在本节中，我会回顾一些在 Web 表格上进行语义标注的相关研究工作，它们通常会处理三个任务：实体链接 (Entity Linking)，列类型推理 (Column Type Inference) 以及关系抽取 (Relation Extraction)。在 Cafarella 等人[12] 报告说，有超过 1.5 亿个 Web 表格内嵌有高质量的关系型数据，许多研究人员意识到 Web 表格是许多应用的重要数据来源，比如信息抽取和结构化数据搜索。因此，与 Web 表格的语义标注相关的各种研究工作如雨后春笋般涌现出来。

Hignette 等人[7] 提出了一种聚合的方法，使用给定本体中的词汇来标注 Web 表格中的内容。它首先标注单元格，然后标注列的类型，最后标注列之间的关系。与之相似的，Syed 等人[10] 提出了一种管道方法，其首先进行列类型推理，然后将单元格的值链接到给定的知识库中的实体，最后选择列之间合适的关系。Zhang [13] 设计了一个名为 TableMiner 的工具来标注 Web 表格。TableMiner 只专注于列类型推理和实体链接，并不能从 Web 表格中抽取关系。之后，Zhang [14] 又提出了一些策略来改进 TableMiner。Limaye 等人[1] 和 Mulwad 等人[8] 提出了两种方法，可以分别对 Web 表格上的实体链接，列类型推理和关系抽取任务联合建模。我的方法和这些研究工作之间的主要区别在于它不依赖任何特定的信息来完成实体链接的任务，例如 Web 表格的列头和标题，知识库中的实体类型和网页中的语义标记等等。

还有一些在特定场景下对 Web 表格进行语义标注的研究工作是没有实体链接的步骤的。在 Venetis 等人[11] 的研究工作中，他们的方法使实体链接的影响削弱，直接进行列类型推理，并且通过大规模的关系数据库 (Relation Databases) 中不同模式的出现频率来确定 Web 表格中的关系，这些关系数据库都是由网页构建的，但通常不对大部分研究人员开放。此外，Munoz 等人[9] 提出了一种从维基百科表格中挖掘 RDF 三元组的方法。在这项研究中，他们能够通过内部链接和文章标题直接识别出维基百科中的实体。

跟我的方法最相近的研究工作是 Bhagavatula 等人[15] 和 Shen 等人[16] 的研究工作。Shen 等人[16] 尝试将形如列表 (List-like) 的 Web 表格 (表格有多行但只有一列) 中的字符串指称链接到给定知识库中的实体。Bhagavatula 等人[15] 提出了 TabEL，一个表格实体链接系统，它使用集体分类技术来对 Web 表格中的所有指称进行联合消岐 (Entity

Disambiguation)。这两个研究工作都不使用任何特定信息来进行实体链接，并且可以应用于任何知识库。在本文中，为了提高 Web 表格中的实体链接的质量，我专注于在多个相互有链接关系的知识库下的实体链接而不是单个知识库下的实体链接。

1.3 本文贡献

之前的 Web 表格实体链接研究中主要存在两个问题。1) 许多研究工作[1][7][8][10][13][14]都非常依赖于基于特定信息的特征，比如 Web 表格的列头 (e.g. “电影”，“导演” 等等。在图 2.1 中表格的第一行中出现)，目标知识库中的实体类型以及其他的一些特定信息。假如我们要处理的 Web 表格中没有这样的列头信息抑或是给定的知识库中没有实体类型的信息，那么很显然前面提及的那些方法的效果会很有限。2) 现在大部分实体链接的方法[1][7][10][13][14][16][15] 都只考虑将 Web 表格单元格中的字符串指称链接到单一知识库，但是每个知识库中的实体数量都是有限的，单一知识库无法保证在做 Web 表格上的实体链接的时候对实体有一个很好的覆盖度。单知识库上的实体链接常常会出现实体缺失的状况。这个问题在这篇论文[17] 中的自然语言文本上实体链接的过程中也有体现。

为了克服上述问题，在我的毕设系统 “基于多知识库的表格实体链接系统” (在后文中都简称 “系统”) 中，我提出了 2 个新的通用的方法来做基于多知识库的 Web 表格实体链接。1) 方法一中包含两个步骤。第一步是将一个不依赖于任何特定信息的基于概率图模型 (Probabilistic Graph Model) 的算法用来做 Web 表格与每个单知识库的实体链接。然后在第二步中，我提出了三个启发式规则，利用来自不同知识库的实体之间的已存在的和新学习的 “sameAs” 关系，以提高第一步的实体链接结果的质量。第二步不仅可以减少单知识库实体链接产生的错误，还可以提高实体链接结果的覆盖范围。2) 方法二基于一种融合的思想，尝试用一个统一的图模型来表示 Web 表格中的字符串指称和来自多知识库的实体信息，然后在这个图模型上进行随机游走 (Random Walk)，直接得到实体链接的结果。简单的说，方法二融合了方法一中的两步，将多知识库间的 “sameAs” 关系加进了概率图模型，同时也舍弃了方法一中的启发式规则，毕竟启发式规则是基于直觉和经验而定，在实际操作中有可能会因为多知识库间 “sameAs” 关系的缺失而舍弃一些非常有价值的实体链接结果，方法二规避了这些风险。在实验中，我将一些 Web 表格样本中的字符串指称与 Zhishi.me[5] 中的实体进行链接，Zhishi.me 是最大的中文链接开放知识库，如图 1.1 所示，其由三个相互链接的中文在线百科知识库组成：中文维基百科[1]，百度百科[2]和互动百科[3]。

针对本文中的 2 个方法，设计了一些对比试验来将它们和目前最先进的实体链接系统 (即 TabEL[15] 和 LIEGE[16]) 进行各方面的比较。实验结果表明，本文中提出的方法在 MRR（即 Mean Reciprocal Rank[4] 平均互惠排名)，准确率，召回率和 F1 值等评价指标上都表现得很不错。实验中用到的表格语料库和整个系统的代码[5]都是公开的，人工标注

[1] https://zh.wikipedia.org
[2] http://baike.baidu.com
[3] http://www.baike.com
[4] https://en.wikipedia.org/wiki/Mean_reciprocal_rank
[5] https://github.com/yanshengjia/link

实体的表格数据同样也对未来的表格实体链接系统开放。

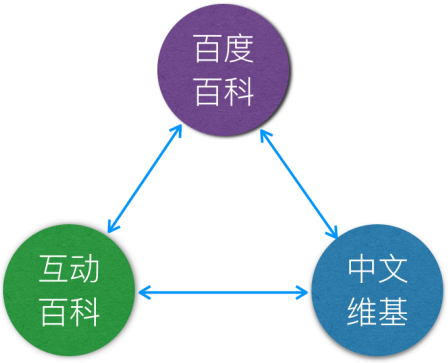

图 1.1 Zhishi.me 由三个相互链接的知识库组成

总而言之，本文的主要贡献在于：

1. 提出了一个两阶段的基于多知识库的表格实体链接方法 (即方法一)，并在实验中体现其比基于单一知识库的实体链接方法的优越性。该方法不依赖表格和知识库中的特定信息，而是建立了一个通用的概率图模型并使用随机游走算法来进行实体的迭代消岐。

2. 提出了一个融合的支持多知识库的表格实体链接方法 (即方法二)，其融合了方法一中的两步，规避了方法一中的启发式规则带来的风险，整个方法都是在一个统一的概率图模型上运行，一步到位得得出链接结果。

3. 设计了一些对比试验，将本文提出的 2 个方法，TabEL[15] 和 LIEGE[16] 在链接准确率、召回率、F1 值和 MRR 值上进行比较，从而验证本文提出的 2 个方法的效果。

4. 在实体消岐时相较于以前的方法使用了很多十分具有价值的特征，比如字符串相似度特征、上下文相似度特征、同义词特征、三元组关系特征、知识库实体的消岐义特征等等。

5. 在系统中添加了一些很有意义的功能，比如实体链接与 “sameAs” 关系的迭代学习。多知识库间的 “sameAs” 关系往往是不完备的，利用实体链接的结果可以发现新的 “sameAs” 关系，利用 “sameAs” 关系也可以来优化实体链接的结果。

本文的各章节内容是这样分配的：第一章是绪论的，介绍了我的毕设项目的研究动机、相关研究工作以及本文的主要贡献；第二章从背景知识的角度出发阐述了基于多知识库的实体链接技术的方方面面，包括任务目标、关键挑战以及一般链接流程；第三章详细介绍了我的毕设系统中的各个模块，包括 2 个方法的思想与其中用到的各种模型与算法；第四章从系统具体实现的角度切入，以一个工程师的身份讲述整个系统的实现细节；第五章是实验与评估，详述了整个实验流程，将本文中的 2 个方法与 TabEL[15] 和 LIEGE[16] 在不同评价指标上进行对比并分析；第六章也就是最后一章总结了全文，并展望了未来。

第二章　基于多知识库的表格实体链接

2.1 任务描述

表格的语义解释的一般任务使用一个表格和一个参考知识库作为输入，通常包括以下三个子任务：

1. 实体链接：找到表格单元格中称为 Mention 的文本短语并将其与对应的知识库参考实体相链接

2. 列类型推理：根据表格中一列包含的实体的知识库类型来推断该列的类型

3. 关系抽取：根据两个表中列与列在给定的一行的实体对的关系将来推断列间关系

实体，类型和关系都是来自在给定的知识库中。举一个具体的例子，给定图 2.1 中的表格和知识库中文维基百科，实体链接的任务就是将字符串指称 “冯小刚” 链接到中文维基百科中的实体**冯小刚**。列类型推理的任务就是将表格第二列与中文维基百科中的导演类型相关联。关系抽取任务则是识别实体**冯小刚**和**非诚勿扰**之间的关系 **isDirectorOf**。在这篇文章中，我专注于第一个语义解释任务，实体链接。实体链接在自然语言处理 (Natural Language Processing) 领域也叫做命名实体消岐 (Named Entity Disambiguation)。需要额外说明的是，我做的是中文的实体链接系统，而不是跨语言的实体链接[18]。接下来会正式定义表格的实体链接任务。还会介绍文章中的用到的一些符号含义。

正式定义

一个表格在系统中被表示为一个矩阵，T，该表格包含 r 行 c 列。使用行和列单位的表格很容易规范化成 $r \times c$ 的矩阵。$T[i, j]$ 代表了 T 在 i^{th} 行和 j^{th} 列的单元格。一个字符串指称 m 指的是表格单元格中的一个字符串，该字符串需要事先被识别出来并且能潜在得指向知识库中的一些实体。在语义网领域中，实体可以理解为独立存在且相互区别的某种事物，并不一定是物质上的存在，可能是一个词语或者一种概念。比如 “迈克尔・乔丹”、“苹果” 都是语义网中的实体。在我的系统中，实体来自最大的中文链接开放知识库 Zhishi.me[5]，其包含了三个相互链接的中文百科型知识库：百度百科、互动百科和中文维基。有一些表格中字符串指称在给定的知识库中可能不存在对应的实体，这样的无对应实体的指称被称为不可链接的指称 (Unlinkable Mention)，在我的系统中会给这样的指称打上一个特殊的标签 “NIL” 来表明它是不可链接的。对于不可链接的指称，现有一些研究[19][20] 会识别它们在知识库中的细粒度类型 (Fine-grained Type)，但这个已

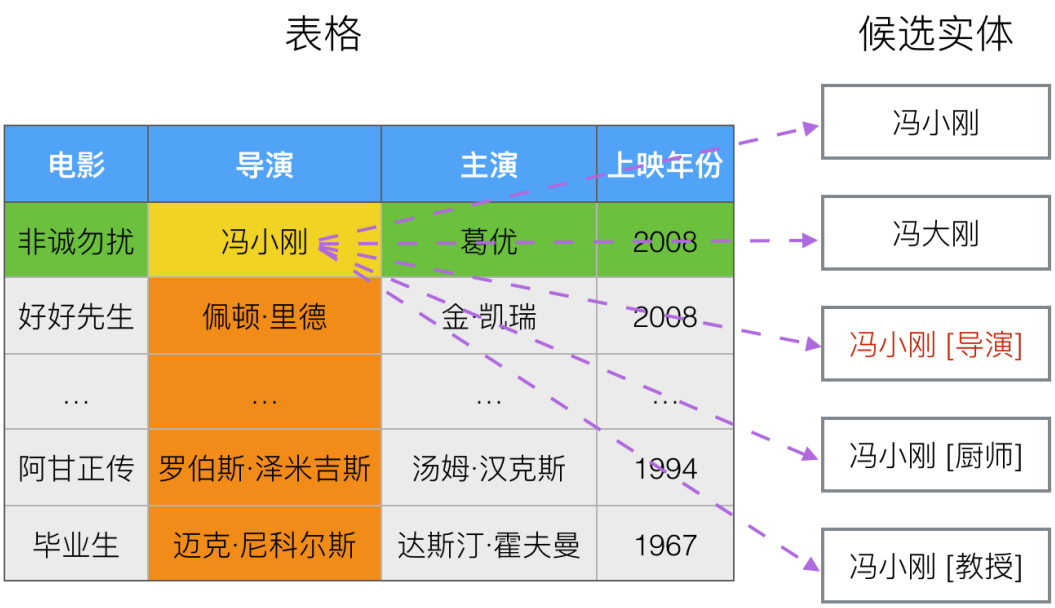

图 2.1 一个对于实体链接任务的演示。左边表格中已识别出的字符串指称位于黄色单元格中。正确的链接实体用红色标出。

经超出实体链接系统的范围。

任务： 给定多个知识库的实体集合 E 和一个表格中字符串指称集合 M，多知识库的表格实体链接的目标就是将表格中的每个字符串指称 $m \in M$ 链接到知识库中它的对应参考实体 $e \in E$。

2.2 关键挑战

现在对实体链接的研究开展得如火如荼，研究者们提出了大量的前景广阔的技术，从深度神经网络到联合推理方法。但是很多论文都没有提及或者思考过实体链接中可能遇到的各项挑战或困难。而这些关键挑战可能会对实体链接的效果产生很大的影响。在实现表格实体链接系统的过程中以及在完成各项对比试验的过程中，我遇到了许多挑战与困难。接下来我会将它们一一罗列出来，希望在未来的工作中能够战胜这些挑战。

1. **缺乏基准数据集**

   在实体链接的各项研究中，一个标准化的基准 (Benchmark) 是非常重要的。这个基准包含 3 个部分：数据集，知识库和算法的评价标准。在大多数研究中，知识库普遍会选取维基百科，算法的评价标准基本上差不多。因此在这里，缺乏基准主要是指缺乏一个标准的高质量的数据集。在英文领域，有一些常用的数据集可用于实体链接算法的训练和测试。比如 UIUC 的 ACE 和 MSNBC 数据集，AIDA 研究组的数据集以及 TAC-KBP 研究组的数据集。但是在中文领域，公开的高质量的实体链接数据集很稀有。这样就导致了研究者在进行实体链接研究之前，需要自己耗费很多时间精力去准备数据集。每个人制作的数据集都是不同的，这也导致了在比较不同研究中的

算法性能时的很多困难。在我的实验中，所有的数据集，也就是 Web 表格，都是从 Web 上人工精心挑选出来的。确保每个表格中的数据准确无误，同时表格中带有许多有歧义的字符串指称，只有这样才能有效地验证算法的消岐性能。

2. **自然语言的多样性和歧义性**
   自然语言在表达上常常带有多样性和歧义性。多样性指的是一义多词，同一意义可以以多种不同的方式表达，同一个知识库实体被多个字符串指称表示。歧义性指的是一词多义，同一个词在不同的上下文中有多种不同的意义，同一个字符串指称可以表示多个不同的实体。自然语言的这 2 个特性都对实体链接带来了一定的挑战，尤其是歧义性。例如，对于字符串指称 “小米”，如果其上下文为 “小明喜欢喝小米粥而不是皮蛋瘦肉粥”，那么它表示的是粮食类实体 “小米”；如果它出现在这样的上下文中，“小米手机真的太酷了”，此时它代表的是手机品牌类实体 “小米”。如果实体链接算法不能很好的理解字符串指称的上下文，那么很有可能链接到的就是一个错误的实体。所以，字符串指称的上下文特征在一些实体链接算法中是相当重要的，能否抽取出高质量的字符串指称上下文特征决定了能否对字符串指称进行正确的消岐。

3. **实体缺失**
   就像 2.1 节中提到的，知识库中的实体数量有限，不可能覆盖世界上的所有实体。因此，在实体链接的过程中，肯定会遇到一些字符串指称没有候选实体，链接不到任何一个知识库实体的情况。这就是实体缺失的问题。在我看来，解决实体缺失最好的办法就是不断扩大知识库的规模，尽可能让它拥有更多实体。现在很多知识库还是依赖于人工撰写页面的方式来收录知识，知识库规模扩大的速度可能永远跟不上 Web 上信息增长的速度，因此实体缺失的问题可能会长期存在。对于 “不可链接 (unlinkable)” 的字符串指称，一些研究者[21][22] 直接给这样的实体打上一个 “NIL” 标签，表示该字符串指称不代表任何实体。预测不可链接实体是实体链接系统的一个重要模块。在我的系统中，为了预测哪些字符串指称是不可链接的，使用了一个简单的启发式的方法，那就是如果一个字符串指称 m 的候选实体集合 E_m 是空集，那么就认为该指称 m 是不可链接的并返回一个 NIL 标签给它。除此之外，还有许多预测不可链接实体的方法，在本文中就不再赘述。

2.3 链接流程

对于一个一般的表格实体链接系统，给定一个表格和一个知识库的实体，系统执行实体链接任务主要分为三步：

1. 指称识别 (Mention Identification)：在表格单元格中识别出每个潜在的指称。

2. 候选实体生成 (Candidate Generation)：对于每个潜在的指称，生成其候选实体集合，知识库中的实体子集都可能是潜在指称的参考实体。

3. 实体消岐 (Entity Disambiguation)：对于每个潜在的指称，根据指称的上下文从其候选实体集合中挑选出一个实体作为字符串指称的参考实体。

我的毕设系统中的 2 个方法基本都遵循这三个步骤，不同之处在于系统的输入变成了多个知识库的实体以及在实体消岐算法部分有一些区别。系统使用的是监督学习的方法，我事先人工标注了表格数据集并使用已标注的指称来训练系统的各个模块。尽管我在实验中使用百度百科、互动百科和中文维基百科作为实验的知识库，但我的方法是通用的，只要给定任意一个知识库中的已标注数据，便可使用该知识库。关于系统中 2 个多知识库实体链接方法的步骤和细节会在第三章中详细介绍。

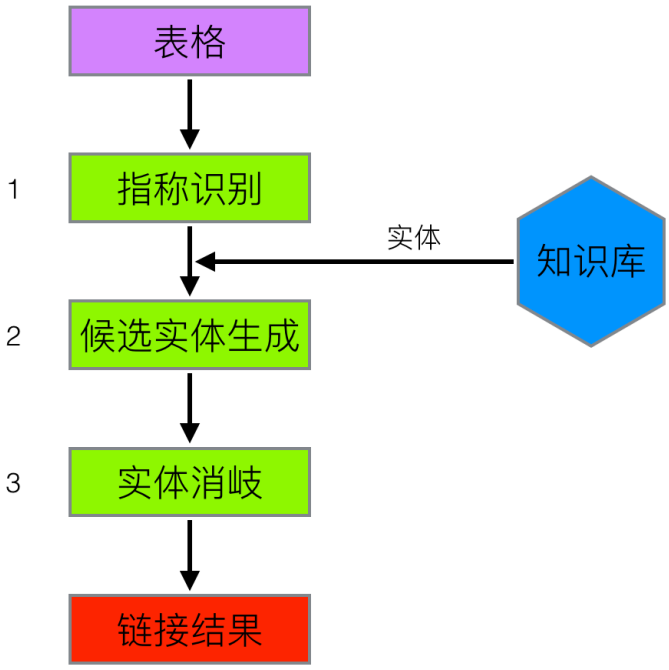

图 2.2 实体链接的一般流程

2.4 本章小结

本章是对多知识库表格实体链接所需的背景知识的一个梗概介绍。先描述了表格语义解释的 3 个子任务，明确了多知识库表格实体链接这个任务的定义，如图 2.1 所示。然后讲述了一些实体链接过程中的关键性挑战，包括我在实验过程中遇到的困难。最后介绍了表格实体链接的一般流程，如图 2.2 所示。在下一章中我会详细得讲述我的毕设系统的设计思路和各模块组成，尤其是 2 个多知识库表格实体链接方法的步骤和异同。

第三章　系统描述

在本章中，我会详细描述系统中使用的概率图模型、随机游走算法以及 2 个方法的细节。这 2 个方法是平行的，方法一是一个两阶段的方法，方法二是方法一的改版，其融合了方法一中的两步，这两个方法的输入和输出都是一样的。输入都是 Web 表格和多知识库的实体数据，输出都是表格的实体链接结果，即表格中的字符串指称最终链接到的知识库中对应的参考实体。换句话说，这 2 个方法中的任何一个都可以单独拿出来作为一个表格实体链接系统的核心算法。我将这 2 个方法放在同一个系统中是为了可以更好地比较二者的效果差异。在使用这个系统时，可以任意选择一个方法，然后得到已选择方法计算得到的实体链接结果。目前实体链接的方法大体上可以分为 3 类：基于概率统计的方法，基于机器学习的方法和基于图模型的随机游走方法。系统中的 2 个方法都是属于基于图模型的随机游走方法。这类方法的思路与另外 2 类方法完全不同。它主要利用字符串指称与实体之间、实体与实体之间的语义相关性来开展实体链接的工作。它认为在位置上相邻的字符串指称往往具有语义相关性，比如表格中同行或者同列的指称描述的一般是同一个事物。在我的方法中，会将字符串指称和知识库实体建模成一张概率图 (Probabilistic Graph Model)，称之为实体消岐图 (Entity Disambiguation Graph)，然后在这张图上运行随机游走 (Random Walk) 算法进行迭代消岐，直到图上实体节点的概率值收敛，最终得到链接结果。

3.1 概率图模型

3.2 随机游走

3.3 方法一: 两步走

方法一包含 2 个主要的步骤：首选使用各个单知识库进行实体链接，然后运行多知识库间的 “sameAs” 关系来优化单知识库的链接结果。

3.3.1 单知识库表格实体链接

指称识别

任何实体链接系统的第一步是识别出潜在的字符串指称，它们能够被链接到知识库中的参考实体。给定来自输入的表格中每个单元格的文本内容，t_q，系统将 t_q 中满足一定条件的最长的短语 s 识别为潜在的指称。这个条件就是对于某些实体 e，字符串 s 能链接

图 3.1 一个表格同行同列中的指称具有语义相关性的例子

到该实体的概率 $P(e|s)$ 非零。如果 s 的长度小于 t_q 的长度，系统会在 s 之后发现长度最长的短语，并以此类推。例如，对于一个单元格的文本“习近平 & 彭丽媛”，系统会识别出到两个潜在的指称：一个是“习近平”，另一个是“彭丽媛”。

候选实体生成

对于表格单元格中的每个字符串指称，首先需要从给定的海量的知识库实体中找出一些可能成为该指称参考实体的实体，来缩小实体链接的范围。这样的实体称为字符串指称的候选实体。这样的过程叫做生成候选实体。在系统中，我讲每个指称分割到单词级别，所以每个指称能被表示为一个单词集合。如果给定知识库中的一个实体 e 或者 s 在 BabelNet[23] (一个全网域多语种同义词辞典) 中的一个同义词包含某个指称 m 的分割单词集合中的至少一个单词，那么实体 e 就被认为是指称 m 的一个候选参考实体。举个例子，字符串指称“苹果”有这样的一些候选实体：“苹果”，“苹果派”，“苹果 [水果]”，“苹果 [智能手机品牌]”。候选实体生成的结果就是每个指称都可能指向一个候选实体集合。在实际操作过程中，除了指称与实体的包含关系，我还考虑了二者之间的字符串相似度 (计算公式在后面会提到)，设置了一个字符串相似度的阈值。一般来说，与指称的字符串相似度很低的实体，很有可能表示的是跟指称完全不同的事物，即便它们有包含关系。所以如果实体 e 和指称 m 的字符串相似度低于阈值，即使 e 包含 m，也不将该实体 e 添加进 m 的候选实体集合。比如，对于指称“苹果”，在知识库中有这样的一个实体“苹果红蜘蛛”，显然二者不可能相链接，虽然这个实体里包含了“苹果”二字，但是由于二者的字符串相似度太低，这样的实体就被剔除了。

实体消岐

3.3.2 多知识库优化实体链接

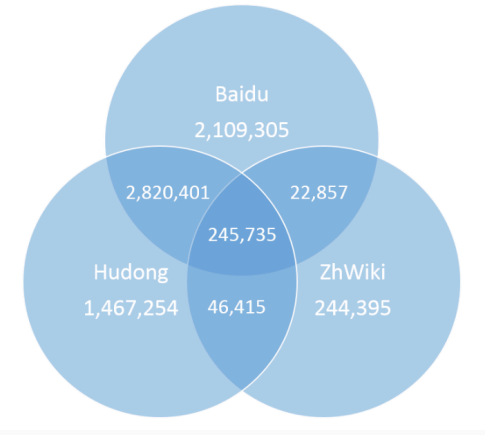

图 3.2 Zhishi.me 数据统计

3.4 方法二: 融合

3.5 EL 与 sameAs 迭代学习

3.6 本章小结

第四章　系统实现

4.1 表格语料库

4.2 知识库实体

4.3 人工标注指称来源

4.4 本章小结

第五章　实验与评估

5.1 本章小结

第六章　总结与展望

6.1 工作总结

6.2 未来展望

致　谢

感谢……

参考文献

[1] Limaye G, Sarawagi S, Chakrabarti S. Annotating and searching web tables using entities, types and relationships[J]. PVLDB, 2010, 3(1-2):1338–1347.

[2] Suchanek F M, Kasneci G, Weikum G. Yago: a core of semantic knowledge[C]. In: WWW. 2007. 697–706.

[3] Auer S, Bizer C, Kobilarov G, et al. Dbpedia: A nucleus for a web of open data[C]. In: ISWC. 2007. 722–735.

[4] Bollacker K, Evans C, Paritosh P, et al. Freebase: a collaboratively created graph database for structuring human knowledge[C]. In: SIGMOD. 2008. 1247–1250.

[5] Niu X, Sun X, Wang H, et al. Zhishi. me-weaving chinese linking open data[C]. In: ISWC. 2011. 205–220.

[6] T. Berners-Lee O L, J. Hendler. The semantic web[J]. Scientific American, 2001.

[7] Hignette G, Buche P, Dibie-Barthélemy J, et al. Fuzzy annotation of web data tables driven by a domain ontology[C]. In: ESWC. 2009. 638–653.

[8] Mulwad V, Finin T, Joshi A. Semantic message passing for generating linked data from tables[C]. In: ISWC. 2013. 363–378.

[9] Muñoz E, Hogan A, Mileo A. Using linked data to mine RDF from wikipedia's tables[C]. In: WSDM. 2014. 533–542.

[10] Syed Z, Finin T, Mulwad V, et al. Exploiting a web of semantic data for interpreting tables[C]. In: WebSci. 2010. 5.

[11] Venetis P, Halevy A, Madhavan J, et al. Recovering semantics of tables on the web[J]. PVLDB, 2011, 4(9):528–538.

[12] Cafarella M J, Halevy A, Wang D Z, et al. Webtables: exploring the power of tables on the web[J]. PVLDB, 2008, 1(1):538–549.

[13] Zhang Z. Towards efficient and effective semantic table interpretation[C]. In: ISWC. 2014. 487–502.

[14] Zhang Z. Learning with partial data for semantic table interpretation[C]. In: EKAW. 2014. 607–618.

[15] Bhagavatula C S, Noraset T, Downey D. TabEL: Entity Linking in Web Tables[C]. In: ISWC. 2015. 425–441.

[16] Shen W, Wang J, Luo P, et al. LIEGE:: link entities in web lists with knowledge base[C]. In: SIGKDD. 2012. 1424–1432.

[17] Pereira B. Entity linking with multiple knowledge bases: An ontology modularization approach[C]. In: ISWC. 2014. 513–520.

[18] T. Zhang J Z, K. Liu. Wu C, Wang H, Qu J, Yu Y[C]. In: IJCAI. 2013. 2218—2224.

[19] W. Shen P L M W, J. Wang. A graph-based approach for ontology population with named entities[C]. In: CIKM. 2012. 345—354.

[20] X. Ling D S W. Fine-grained entity recognition[C]. In: AAAI. 2012.

[21] T. Zhang J Z, K. Liu. ZhishiLink: Entity Linking on Zhishi.me[C]. In: Proceedings of the 7th CSWS. 2013. 161–174.

[22] Dredze M R D G A F T, McNamee P. Entity Disambiguation for Knowledge Base Population[C]. In: Proceedings of the 23rd International Conference on Computational Linguistics. 2010. 277–285.

[23] Navigli R, Ponzetto S P. BabelNet: Building a very large multilingual semantic network[C]. In: ACL. 2010. 216–225.

附录 A　第一个附录

……

作者简介 (包括论文和成果清单)

作者简介