摘 要

词嵌入是近年来比较热门的一个研究方向，词嵌入的目的是通过低纬度的稠密向量表征出知识库中的实体和关系的语义信息，在有效存储信息的基础上，能够实现实体预测、链接预测等推理任务。目前，大多数的工作集中在通过优化评分函数、损失函数等来训练出一个更有效的翻译模型，使得训练出的翻译模型可以在meanrank、Hit@10的评分标准下取得更好的成绩。本文考虑到翻译模型训练数据与本体语义的相关性，利用本体推理研究的现有成果提出一种新的方法来提升构建出的翻译模型的性能。这种方法使用本体完成对训练集的语义表达，同时使用测试数据中的三元组集合构建出观察值的集合。通过这种方式，本文把优化翻译模型的问题，转化成本体研究中的溯因推理问题。溯因诊断是本体新知识获取的一个重要的推理方法，这种方法可以揭示出观察值没有被成功推理出的根本原因，并且可以提供出与原本体语义一致的解释集。解释集中的公理与本体中的公理具有相同的语法，在使用获得的解释集对缺陷本体进行完善修复后，新的本体不仅可以蕴含观察值，还因为关键性公理的补全，有着更好的获取新知识的能力。在本体的溯因诊断中，限制解释集的大小是必须考虑的一个关键因素。解释集过于强大的表达能力会导致解空间的无限膨胀，而过度地压缩解的空间，不仅会极大地削弱解释集的表达能力，也会影响解释集的对缺漏公理的覆盖能力。为了平衡解释集的表达能力与解空间的矛盾，本文根据决断集合的特点提出一个新的概念，决断集合模版。基本地，决断集合模版由决断集中获取，获取方法是把决断集中的公理进行变量化，解释集会从原本体以及决断集模版的集合中进行计算，然后实例化得来。同时，为了限制解释集的空间，本文会介绍一个新的概念，最小决断集模版。利用最小决断集模版，我们可以有效地限制解释集的空间大小。在本体的溯因诊断中，观察值的表达是另外一个需要考虑的因素，但是在词向量模型的训练集的数据中，三元组是它的唯一表达方式，因此本文将会使用简化的表达语言完成对观察值集合的表达，这样可以使得解释集的解空间缩小而不影响解释集的表达能力。本文的目标是通过利用本体的表达能力对词嵌入的训练数据进行表达，从而把优化词向量模型的问题转化成本体研究中的溯因诊断问题。通过利用决断集模版的语义能力，我们可以从模版和集合中实例化出目标解释集。然后，利用诊断出来的结果，我们可以对词向量的训练数据进行修正，从而达到优化训练出来的词向量模型的目的。最后，我们还设置了一组相关的实验，验证本文得出的解释集对训练数据的修正能力，以及对比不同方法下所得出

的词向量模型的表达能力和推理能力。

关键词：本体；词向量；溯因诊断

Abstract

Word embedding is a popular research direction in recent years. The purpose of word embedding is to express the semantic information of entities and relationships in the knowledge base through the dense vector of low latitude. Based on the effective storage of information, it can perform entity prediction, link Prediction and other reasoning tasks. At present, most of the work focuses on training a more efficient translation model by optimizing the scoring function, loss function, etc., so that the trained translation model can achieve better results under the score of meanrank and Hit@10. This paper takes into account the relevance of the translation model training data to the ontology semantics, and puts forward a new method to enhance the performance of the translated model by using the existing results of ontology reasoning research. This method uses the ontology to complete the semantic expression of the training set, and uses the set of triples in the test data to construct a set of observations. In this way, this paper transforms the problem of optimizing the translation model into the problem of abductive reasoning in ontology research. The retrospective diagnosis is an important reasoning method for the acquisition of new knowledge of ontology. This method can reveal the root cause that the observed value has not been successfully entailed, and can provide the explanation which is consistent with the original body semantics. The axiom of the explanation has the same grammar as the axiom in the ontology. After using the obtained explanation to perfect the defect ontology, the new ontology can not only contain the observation value, but also because of the completeness of the key axiom, The ability to acquire new knowledge. In the ontology diagnosis of the ontology, limiting the size of the explanatory set is a key factor that must be considered. Interpretation of the set of too strong expression will lead to the infinite expansion of the solution space, and excessive compression of the solution space, not only will greatly weaken the interpretation of the set of expression ability, will also affect the interpretation of the lack of coverage of the lack of ability. In order to balance the contradiction between the expression ability of the explanation and the solution space, this paper proposes a new concept based on the characteristics of the de-

cision set. Basically, the decision set template is obtained from the decision, and the acquisition method is to quantify the axioms of the decision set, and the explanation assembly is calculated from the original set and the set of decision sets, and then instantiated. At the same time, in order to limit the space of the interpretation set, this article will introduce a new concept, the smallest template. In the ontology diagnosis of the ontology, the expression of the observation is another factor to consider, but in the data set of the training set of the word vector model, the triplet is its only expression, so this article will use the simplified expression The language completes the expression of the set of observations, which allows the solution space of the interpretation set to be reduced without affecting the expression of the interpretation set. The goal of this paper is to express the training data of the word by using the expression ability of the ontology, so as to transform the problem of the optimization word vector model into the dying diagnosis problem in the ontology research. By using the semantic capabilities of the decision set template, we can extract the target set from the template and the set of examples. Then, using the results of the diagnosis, we can modify the training data of the word vector so as to achieve the purpose of optimizing the vector model of the training. Finally, we also set up a set of related experiments to verify the corrective ability of the explanatory set to be derived from the training data, and to compare the expression ability and reasoning ability of the word vector model obtained under different methods.

Keywords: Ontology; Word Vector; Abduction Diagnosis

目 录

第一章　引言

目前本体并没有统一的定义和固定的应用领域，斯坦福大学的Gruber给出的定义得到了许多同行的认可，即本体是共享概念模型的明确的形式化规范说明[?]。本体提出的最初目标是实现知识的共享、集成和重用。一个完善的本体可以澄清领域知识的结构，通过构建一个统一结构或者一个规范模型来减少概念和术语上的差异。同时，利用本体对需求解决的问题和任务进行规范化描述，可以提高需求分析、信息获取的效率，节约成本[?]。

1.1　本文的意义

进入信息时代，人们花费了大量的精力构建结构化的知识库。利用已有的知识库，进行表示学习的目标，是通过转化规则将现有的知识库的语义信息转化到低维的向量中进行表示，并利用转化后得到的低维向量学习出新的知识。但是，构建出的向量模型与现实世界观察值的不一致是一种出现频率较高的问题，这种不一致会降低向量模型推理结果的可信度，同时也意味着已构建的向量模型对语义信息表达的准确度有待提高。因此，如何找到产生这种不一致的原因，以及如何对已构建的向量模型进行修复，是目前研究的重点。然而，目前的研究工作，大多数集中在改进向量模型生成的算法上，但是对由知识库产生的问题，却还没有做较为深入的研究。知识库是向量模型的生成来源，一个不完整的或者是有缺陷的知识库，会对生成的模型产生重大的影响。与此同时，对本体的研究获得了较大的进展。本体作为语义网的核心，是一种清晰表达语义和知识共享的方式，也就是在特定领域之中那些存在着的对象类型或概念及其属性和相互关系，在本体中进行推理可以获得知识库中没有进行表达的知识，而其中，溯因诊断又是本体中一种重要的推理方式。溯因诊断是推理到最佳解释的过程，它是开始于事实的集合并推导出它们的最合适的解释的推理过程，溯因意味着生成假设来解释观察货结论。本文的主要目标，是利用本体语言对知识库与观察值进行表达，再在生成的本体中进行溯因诊断，利用针对词向量知识库改进的诊断方法，找出知识

库产生不一致的原因，并在此基础上，实现对知识库的修复，从而达到修正向量模型的目的。

1.2 研究现状

本体学习最早是由Alexander Madche和Steffen Staab 两人提出的，他们把本体学习描述成从数据中提取领域模型[?]。本体学习是一个广泛的概念，涉及的领域和学科众多，目前研究者们把本体学习分割成以下几个部分。在大多数情况下，对于每一层任务的研究都是建立在更底层的研究基础之上。目前对本体学习的研究集中在基础五层任务之中，包括术语提取、同义词识别、概念提取、概念层次关系提取、非分类关系的提取。

目前本体学习主要有两种分类方法，分别是按照信息输入源分类和按照构建的方法进行分类。

现阶段本体学习的信息输入源有三类，分别是基于结构化数据的本体学习，基于半结构化数据的本体学习和基于非结构化数据的。结构化数据主要包括关系数据库或面向对象数据库中的数据，半结构化数据是指具有隐含结构，但缺乏固定或严格结构的数据，目前web上也有不少半结构化数据，以RDF标注的网页也越来越多[?]。非结构化数据指的是纯文本无结构化标注的自然语言数据，文章、访谈甚至剧本都可以看做是本体学习的非结构化信息源。目前Web上数量最大、信息最多的就是这一类型的非结构化数据，同时这一类型数据也是价值含量最高的，所以目前的研究多集中于以非结构化数据作为信息源的本体学习[?]。

现有的本体构建方法有手工和自动两种。手工构建指的是有由领域的专家参与进行本体构建工作，本体的完全手工构建耗时、费力，容易出现倾向性错误，且难以及时地更新。另一种本体构建的方法是自动构建法，通过结合自然语言处理、机器学习、人工智能等多领域的研究成果在给定的信息输入源中自动构建出本体。自动构建法分为两种方式，一种是全自动，全自动方指的是在输入源的基础上构建出一个全新的本体。另一种是半自动的方式，半自动构建本体意味着在输入信息源的基础上还有一个该领域的初步本体模型，半自动构建本体的任务就

是从信息源中挖掘出更多的概念和关系用以扩充或者纠正原有的初步本体模型。本体的自动构建具有相当大的潜在价值，但是由于自动构建的难度较大，设计的学科领域众多，现时关于本体自动学习的研究尚未达到完全成熟的水平。[**?**]

本体的手工构建繁琐且难度较大，因此现有的手工构建的本体并不多，其中以普林斯顿大学认识科学实验室在心理学教授乔治?A?米勒的指导下建立和维护的英语字典wordnet最为出名。有别于普通意义上的字典，Wordnet包含了语义信息。Wordnet根据词条的意义将它们分组，每一个具有相近意义的词条组成了一个synset(同义词集)。Wordnet为synset提供了简短，概要的定义，并记录不同synset之间的语义关系。Wordnet发展到现在，已经具有了一定的完整性，涵盖额大多数的英语单词，缺点就是目前尚未支持英语以外的其它语言目前本体的自动构建并没有统一的步骤，得到大多数人认可的是由本体构建工具Protégé开发商斯坦福大学医学院生物信息研究中心提出的七步法[**?**]。七个步骤分别是：(1) 确定本体的专业和范畴

(2) 考察服用现有本体的可能性

(3) 列出本体中得重要术语

(4) 定义类和类的等级体系(完善等级体系可行的方法有：自顶向下法、自底向上法和综合法)

(5) 定义类的属性

(6) 定义属性的分面

(7) 创建实例

斯坦福大学医学院提出的七步法步骤清晰，合理，且经过实验验证在不同领域都有具有较好的适用性。在进行本体构建的时候并不一定要严格遵循以上七个步骤，可根据构建目标领域的需求进行调整以及适应性更改。以自然语言文本作为信息源的需要对文本进行预处理，在这方面自然语言处理的研究已经比较完善，由斯坦福大学开发的Standford Parser取得了不错的成绩。在中文处理方面，由国内复旦大学开发的复旦自然语言处理开发套件可以以较高的准确度完成分词、实体识别、依存句法树生成等任务。在术语提取方面，最主要的方法是由G.

Salton和C. Buckley提出的名词索引的方法[**?**]。他们通过统计学的算法计算出文档中每个词的权重，通过设定阀值筛选出推荐的候选术语。同时有一部分的研究专注于以自然语言处理的方法提取术语，缺点是提取的精度受自然语言处理水平的影响较大。本体学习的另一项重要的任务是从信息源中提取出概念和概念之间的关系。大多数对概念之间关系的提取都使用了数据分析法以及或多或少的语义分析相关的方法。其中最具有代表性的是由D.Faure提出的以动词为核心的概念关系提取，他把实体的属性看成是概念之间的关系。缺点是以动词为核心的提取方法只能提取非分类关系，对于part-of,等层次关系提取成功率较低。

1.3 本文的工作

本文的主要工作是结合国内外已有的本体学习相关的研究成果，针对中文语法特点和语言习惯，从给定的领域语料库中构建出本体，并映射到以RDFS表示的模型当中，具体的步骤如下：(1) 利用复旦大学开发的fnlp中文自然语言处理开发库对输入的中文自然语言的领域语料库进行预处理，包括分词处理，实体名识别，依存句法树生成等。

(2) 从生成的依存句法树中提取出主谓宾结构，转化成知识三元组结构，构建出初步的RDF模型。

(3) 结合统计方法和与语义分析方法，在给定的语料库抽取出领域术语集。

(4) 分析出领域信息源中的概念，并在wordnet的帮助下识别出领域中的上下位关系(即RDFS中得子类关系)

(5) 利用获得的初步RDF模型和提取出来的领域术语及其关系生成该领域本体的RDFS模型。

1.4 论文结构简介

本文主要分五部分。第一部分是对待解决的问题进行简单的描述，描述了本体学习现在的研究进度，并对各个方面的研究水平给出评价，并概括全文的工作。第二部分是预备知识，简单阐释了了本文工作所需的基本知识。第三部分是

本文的核心部分，重点讲述了本文从中文自然语言文本构建出本体的详细方法。第四部分是实验部分，实验部分是本文对理论的验证。最后是致谢部分，以此感谢对写作本文予以帮助的人。

第二章　预备知识

2.1　知识表示学习

2.2　本体语言以及推理

2.2.1　本体与一致性

对于描述逻辑语言$\mathcal{L}$包含以下三个集合，分别是一个由个体名组成的集合N_I，一个由概念名组成集合N_C以及一个由二元关系名组成的集合N_R。我们把$\mathcal{L}$ 的语义演绎表示为$\mathcal{I} = (\Delta^{\mathcal{I}}, \cdot^{\mathcal{I}})$，其中$\Delta^{\mathcal{I}}$表示$\mathcal{I}$的定义域，是一个非空的个体集合，$\cdot^{\mathcal{I}}$是一个映射函数，这个映射函数可以完成概念名到$\Delta^{\mathcal{I}}$子集的映射，二元关系到$\Delta^{\mathcal{I}} \times \Delta^{\mathcal{I}}$ 子集的映射以及个体到$\Delta^{\mathcal{I}}$元素的映射。在本文中，我们考虑表2.1中的语言片段。

表 2.1: 语法以及语义表

Constructor	Syntax	Semantics
top concept	$\top$	$\triangle^{\mathcal{I}}$
bottom concept	$\bot$	$\emptyset$
conjunction	C $\sqcap$ D	$C^{\mathcal{I}} \cap D^{\mathcal{I}}$
existential restriction	$\exists r.C$	$\{X \in \triangle^{\mathcal{I}} \mid \exists y \in \triangle^{\mathcal{I}}: (x, y) \in r^{\mathcal{I}} \wedge r \in C^{\mathcal{I}}\}$
general concept inclusion	$C \sqsubseteq D$	$C^{\mathcal{I}} \subseteq D^{\mathcal{I}}$
role inclusion	$r_1 \circ \cdots \circ r_k \subseteq r$	$r_1^{\mathcal{I}} \circ \cdots \circ r_k^{\mathcal{I}} \subseteq r^{\mathcal{I}}$

一个DL本体$\mathcal{O} = (\mathcal{T}, \mathcal{A})$包括两个部分，一个是TBox $\mathcal{T}$，它描述术语知识与应用领域相关的背景知识，处理概念的定义，由有限个公理构成，其中有引入新概念名和角色名称的公理，有断言包含关系的公理以及断言觉得可传递角色或功能性角色公理[面向Web 的个性化语义信息检索技术]。另一个是断言知识的集合ABox $\mathcal{A}$，它描述的是TBox词汇表中的个体断言,包括概念类的成员元素，二元关系的成员元素二元组以及二元关系的等价关系。在本文中，断言知识部分我们仅考虑二元关系的断言，形如$r(a, b)$，其中r是$\Delta^{\mathcal{I}} \times \Delta^{\mathcal{I}}$ 中的一个二元关系实例，a和b是$\Delta^{\mathcal{I}}$中的一个个体。表2.2展示了TBox与ABox的语法以及语义。

表 2.2: DL示例公理的语义

Syntax	Semantics
$C \sqsubseteq D$	$C^{\mathcal{I}} \subseteq D^{\mathcal{I}}$
$C \equiv D$	$C^{\mathcal{I}} = D^{\mathcal{I}}$
$C(a)$	$a^{\mathcal{I}} \in C^{\mathcal{I}}$
$r(a,b)$	$\langle a^{\mathcal{I}}, b^{\mathcal{I}} \rangle \in r^{\mathcal{I}}$

一个$\cdot^{\mathcal{I}}$如果能够满足本体$\mathcal{O}$中的所有公理，那么这个$\cdot^{\mathcal{I}}$就被称为本体的一个模。因此我们有以下定义：

定义 2.1 (一致) 对于一个DL本体$\mathcal{O}$，如果它有至少一个模，那么这个本体就会被称为是一致的，记作$\mathcal{O} \nvDash \bot$。相反地，如果一个DL本体$\mathcal{O}$没有至少一个模，那么这个本体就会被称为不一致的，记作$\mathcal{O} \vDash \bot$。

2.2.2 本体决断集

本体包含TBox与ABox，因此本体具有从已有的知识库中获取新知识的能力，这种本体中没有，但是可以通过推理出来得到的公理就叫做蕴含。一个本体可能会有一个或多个蕴含，在本体的开发中，能够从蕴含逆推出在推理过程相关的公理具有重大的现实意义，这些公理的集合我们称之为决断集。

定义 2.2 (决断集) 令$\mathcal{O}$为一个一致的DL本体，且$\mathcal{O} \vDash \alpha$，其中$\alpha$是蕴含。对于$\mathcal{O}$的一个子集$\mathcal{O}'$，如果对于$\mathcal{O}'$的所有子集$\mathcal{O}''$满足$\mathcal{O}'' \nvDash \alpha$且$\mathcal{O}' \vDash \alpha$，那么$\mathcal{O}'$就是本体$\mathcal{O}$ 中对于α的一个决断集。

例 2.1 考虑一个一致的DL本体$\mathcal{O}$，其中TBox包含以下三条公理:

(1) $\mathsf{Girl} \sqsubseteq \mathsf{Female}$

(2) $\mathsf{Female} \sqsubseteq \mathsf{Person}$

(3) $\exists \mathsf{giveBirth}.\mathsf{Person} \sqsubseteq \mathsf{Female}$

ABox包含以下两条公理:

(1) Female$(Mary)$

(2) giveBirth$(Lily, Mary)$

蕴含为：

Female$(Lily)$

2.3 描述逻辑溯因诊断

2.3.1 论域词汇表与术语断言

2.3.2 溯因诊断

逻辑研究的是基于规则的推理方式，目前的研究中把推理的方式分为三类，分别是演绎、归纳和溯因推理。演绎推理是最常用的推理方式，演绎推理根据已有的前提事实以及规则，得出结论。对于相同的输入，如果严格按照规则进行运算，演绎推理具有相同的输出，具有恒真性(truth-preserving)。归纳推理则是在已知事实的集合中寻找共同特性，推导出更多事实或同类事实的性质[论语用推理的逻辑属性]。它的推理格式形如以下形式：

a. 所有已知的A为B。

b. 因此，A为B。

溯因推理是推理方式中的第三种方式，溯因推理的方式与前两种推理方式有着本质的区别，溯因推理又称作反绎推理，是推理到最佳解释的过程。一般的，它是开始于事实的集合并推导出它们的最合适的解释的推理过程。术语溯因（abduction）意味生成假设来解释观察或结论。因为需要生成假设来解释观察或结论，因此溯因推理会在进行解释的过程中为前提事实增加新的知识使得前提事

实与解释的并集可以通过演绎推理的方式演绎出观察值或结论。在描述逻辑本体中，溯因推理是一种重要的推理方式。在描述逻辑本体的构建过程中，本体由于构建不够完善，会经常性出现本体无法蕴含观察值的现象。因此，这个时候就需要进行利用溯因推理的方法，已构建的本体进行诊断，找出本体不完善的原因，并在找到的原因的基础上，提出解释对本体进行修复。这种找出原因并提出解释的推理方法就叫做溯因诊断。溯因诊断对修复本体有着重要的意义。我们对术语断言的溯因诊断问题进行了如下两个定义[ABox Abduction in the Description Logic ALC]：

定义 2.3 (术语断言溯因诊断问题) 令$\mathcal{L}_K$和$\mathcal{L}_Q$为两个DL本体，$\mathcal{K} = (\mathcal{T}, \mathcal{A})$是一个$\mathcal{L}_K$中的一个知识库，$\Phi$是一个在$\mathcal{L}_Q$中的术语断言集合。当且仅当$\mathcal{K} \nvDash \Phi$且$\mathcal{K} \cup \Phi \nvDash \bot$的时候，二元组$\langle \mathcal{K}, \Phi \rangle$被称作术语断言的溯因诊断问题。

定义 2.4 (术语断言溯因诊断解释) 令$\mathcal{L}_S$为一个DL本体，且A为一个多个在$\mathcal{L}_S$中的术语断言集合。对于一个溯因诊断问题$\langle \mathcal{K}, \Phi \rangle$，当且仅当$\mathcal{K} \cup A \vDash \Phi$我们把$A$称为可接受解释。更多的，我们把$A$称为：

1. (一致) 当且仅当$\mathcal{K} \cup A \nvDash \bot$。

2. (非平凡) 当且仅当$\mathcal{K} \nvDash \Phi$。

3. (最小) 当且仅当不存在对于$\langle \mathcal{K}, \Phi \rangle$问题的解释$B$，其中$B$是$A$的实例化子集。我们说$B$是$A$的实例化子集当且仅当存在一个重命名映射$\rho$:$N_I^*(B) \mapsto N_I^*(A)$，其中$N_I^*(B)$和$N_I^*(A)$是来自于$A$和$B$的个体名且不出现于$\mathcal{K}$，使得$A \vDash \rho B$。但是对于所有的$\varrho$:$N_I^*(A) \mapsto N_I^*(B)$满足$B \nvDash \varrho A$。满足以上条件的，我们称$A$是问题$\langle \mathcal{K}, \Phi \rangle$的最小解释。

术语断言的溯因诊断需要计算出由一条或多条术语断言的集合，这些集合需要满足最小集的条件。当这些集合被加入到本体中的时候，它需要保持本体的一致性，并且可以使得更新后的本体能够蕴含(⊨)含观察值。$\mathcal{O} \vDash \alpha$表示对于所有满足本体$\mathcal{O}$的模，都可以使得$\alpha$成立。

2.4 定义

定义 2.5 (可扩展公理) 对于本体的一条公理，如果公理中的一个或多个的二元关系或者个体被二元关系变量或者个体变量替换，则这条公理会被称作可扩展公理。一条可扩展公理会被称为可全扩展公理如果这条公理的所有二元关系和个体都被二元关系和个体变量替换。

定义 2.6 (替换) 对于一条可扩展公理或者一个由可扩展公理构成的集合E，替换是一个从E中的二元关系变量或个体变量到其它二元关系变量或者二元关系以及个体变量或个体的映射。其中，如果该替换把所有二元关系变量映射到二元关系以及所有个体变量映射到个体，我们就把这个替换称作实例化替换。

定义 2.7 (解释) 给定一个一致的本体$\mathcal{O}$以及一个二元关系实例α，$\mathcal{O} \nvDash \alpha$, 并且$\mathcal{O} \cup \{\alpha\}$ 是一致的，那么假如存在一个公理的集合$\mathcal{E}$使得$\mathcal{O} \cup \mathcal{E} \models \alpha$，$\mathcal{E} \nvDash \alpha$并且$\mathcal{O} \cup \mathcal{E}$ 是一致的，那么我们称这个集合$\mathcal{E}$为在本体$\mathcal{O}$中对α的解释。

定义 2.8 ($\subseteq_{\text{ds}}$-*minimal* 解释) 一个解释$\mathcal{E}$会被称为$\subseteq_{\text{ds}}$-*minimal* 如果这个$\mathcal{E}$满足以下条件：不存在这样一个解释$\mathcal{E}'$使得$\mathcal{E}' \subseteq_{\text{ds}} \mathcal{E}$ 且$\mathcal{E} \not\subseteq_{\text{ds}} \mathcal{E}'$，其中$\mathcal{E}' \subseteq_{\text{ds}} \mathcal{E}$ 表示存在一个$\mathcal{E}'$的差异化替换θ，使得$\mathcal{E}'\theta \subseteq \mathcal{E}$。

定义 2.9 (基于决定集模版的解释) 对于一个给定的一致本体$\mathcal{O}$，一个基于$role(X, Y)$的模版$\mathcal{P}$以及一个观察值α其中$\mathcal{T} \nvDash \alpha$且$\mathcal{O} \cup \{\alpha\}$是一致的，对于$\alpha$在$\mathcal{O}$ 的解释$\mathcal{E}$会被称为基于决定集模版的解释，如果这个解释满足以下四个条件：

- (非平凡) $\mathcal{E} \nvDash \alpha$
- (一致) $\mathcal{O} \cup \mathcal{E}$ 是一致的
- ($\subseteq_{\text{ds}}$ $-$*minimal*) 不存在一个对于$alpha$在$\mathcal{O}$中的解释$\mathcal{E}'$使得$\mathcal{E}' \subseteq_{\text{ds}} \mathcal{E}$ 且$\mathcal{E} \not\subseteq_{\text{ds}} \mathcal{E}'$。
- (可容许) 存在一个决定集模版$\mathcal{J}_p \in \mathcal{P}$以及$\mathcal{J}_p$的一个差异化替换$\theta$使得$\alpha = role(X\theta, Y\theta)$，$\mathcal{E} \subseteq \mathcal{J}_p\theta$且$\mathcal{J}_p\theta \in Jst(\alpha, \mathcal{O} \cup \mathcal{E})$

定义 2.10 (诊断问题) 我们把$\mathcal{P} = (\mathcal{T}, \mathcal{A}, \alpha)$称作一个诊断的问题实例，其中本体$\mathcal{O} = \mathcal{T} \cup \mathcal{A}$是一个一致的描述逻辑本体。对于问题$\mathcal{P}$的一个解释$\mathcal{E}$满足：$\mathcal{T} \cup \mathcal{A} \cup \mathcal{E} \models \alpha$ 且$\mathcal{T} \cup \mathcal{A} \cup \mathcal{E} \not\models \bot$。

定义 2.11 (决定集) 对于一个一致的描述逻辑本体$\mathcal{O}$，且$\mathcal{O} \models \alpha$($\alpha$是一个推论)，$\mathcal{O}$的子集$\mathcal{O}'$会被称作在本体$\mathcal{O}$中对于$\alpha$的决定集如果$\mathcal{O}' \models \alpha$，且对于所有的$\mathcal{O}'' \subset \mathcal{O}'$ 都有$\mathcal{O}'' \not\models \alpha$。

定义 2.12 (决定集模版) 一个由扩展公理组成的集合$\mathcal{J}_p$会被称为在本体$\mathcal{O}$中对于$role\,(x, y)$的决定集模版如果$\mathcal{J}_p$满足以下两个条件:(1) 存在一个替换θ使得$\mathcal{J}_p\theta \in \mathit{Jst}\,(role\,(X\theta, Y\theta)\,, \mathcal{O})$ (2)对于所有的实例化替换σ存在$\mathcal{J}_p\sigma \models role\,(X\sigma, Y\sigma)$

命题 2.13 给定一个一致的本体$\mathcal{O}$, 一个在$\mathcal{O}$中对于$role(X, Y)$的决定集模版$\mathcal{P}$，以及一个观察值$role(A, B)$，其中$\mathcal{O} \not\models role(A, B)$ 且$\mathcal{O} \cup \{role(A, B)\}$是一致的，那么对于观察值$role(A, B)$在本体$\mathcal{O}$中由$\mathcal{P}$映射得到的解释集合$\mathcal{S} = \{\mathcal{J}_2\theta \mid \mathcal{J}_p \in \mathcal{P}, (\mathcal{J}_1, \mathcal{J}_2) \in \mathit{bipart}(\mathcal{J}_p)$,且$\theta$是一个在$\mathcal{J}_1 \cup \{role(X, Y)\}$上的替换使得$X\theta = A, Y\theta = B, \mathcal{J}_1\theta \subseteq \mathcal{O}, \mathcal{J}_2\theta \not\models role(A, B)\}$。

命题 2.14 一个对于观察值α在本体$\mathcal{O}$中的$\subseteq_{\mathsf{ds}}$-*minimal* 解释$\mathcal{E}$同时也是一个*subset-minimal*的解释。

命题 2.15 给定一个一致的本体$\mathcal{O}$, 一个在$\mathcal{O}$中对于$role(X, Y)$的决定集模版$\mathcal{P}$，以及一个观察值$role(A, B)$，其中$\mathcal{O} \not\models role(A, B)$ 且$\mathcal{O} \cup \{role(A, B)\}$是一致的，那么对于观察值$role(A, B)$在本体$\mathcal{O}$中由$\mathcal{P}$映射得到的解释集合$\mathcal{S} = \{\mathcal{J}_2\theta \mid \mathcal{J}_p \in \mathcal{P}, (\mathcal{J}_1, \mathcal{J}_2) \in \mathit{bipart}(\mathcal{J}_p)$,且$\theta$是一个在$\mathcal{J}_1 \cup \{role(X, Y)\}$上的替换使得$X\theta = A, Y\theta = B, \mathcal{J}_1\theta \subseteq \mathcal{O}, \mathcal{J}_2\theta \not\models role(A, B)\}$。

例 2.2 (诊断) 考虑以下训练集：

1) $(Mike, /isFatherOf, Josan)$

2) $(Mike, /isParentOf, Lily)$

3) $(Hill, /isFriendOf, Peter)$

关系路径: 观察值:

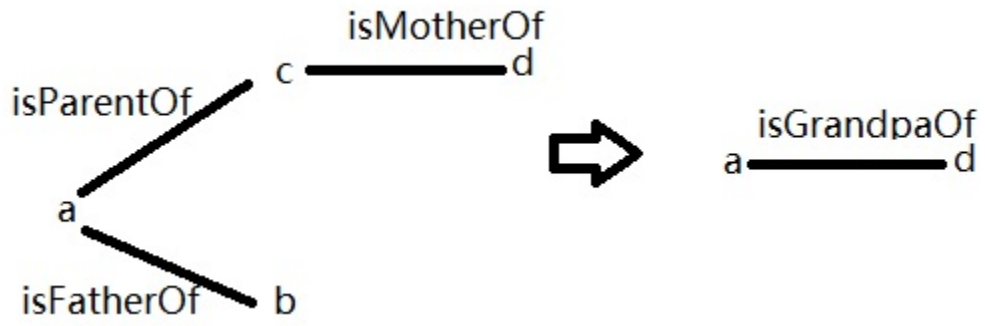

图 2.1: 关系路径

$$(Mike, /isGrandpaOf, Peter)$$

对于例中的训练集，关系路径以及观察值，我们使用本体语言对其进行表达，分别得到TBox $\mathcal{T}$,ABox $\mathcal{A}$以及观察值α。

其中$\mathcal{T}$为：

$$isFatherOf\,(a,b) \wedge isParentOf\,(a,c) \wedge isMotherOf\,(c,d) \rightarrow isGrandpaOf\,(a,d)$$

$\mathcal{A}$为:

1) *isFatherOf*$(Mike, Josan)$

2) *isParentOf*$(Mike, Lily)$

3) *isFriendOf*$(Hill, Peter)$

观察值α为:

$$\textit{isGrandpaOf}(Mike, Peter)$$

使用二分法对$\mathcal{T}$中的公理进行划分，分别得到以下公理集：

1) J_1: $isFatherOf\,(a,b) \wedge isParentOf\,(a,c) J_2$: $isMotherOf\,(c,d)$

2) J_1: $isParentOf\,(a,c) \wedge isMotherOf\,(c,d) J_2$: $isFatherOf\,(a,b)$

3) J_1: $isFatherOf\,(a,b) \wedge isMotherOf\,(c,d) J_2$: $isParentOf\,(a,c)$

4) J_1: $isMotherOf\,(c,d) J_2$: $isFatherOf\,(a,b) \wedge isParentOf\,(a,c)$

5) J_1: $isFatherOf\,(a,b) J_2$: $isParentOf\,(a,c) \wedge isMotherOf\,(c,d)$

6) J_1: $isParentOf\,(a,c) J_2$: $isFatherOf\,(a,b) \wedge isMotherOf\,(c,d)$

首先，我们使用第1个公理集中的J_1 在$\mathcal{A}$中进行实例化，得出以下替换$\mathcal{M}$:

$$\{a \to Mike, b \to Josan, c \to Lily, d \to Peter\}$$

根据得出的替换$\mathcal{M}$,对公理集中的J_2执行实例化，因此得到的第一个解释$\mathcal{E}_1$为：

$$\{\textit{isMotherOf}(Lily, Peter)\}$$

句法分析是自然语言处理领域的一个关键问题，正确地解析出句法的结构对于自然语言处理的相关任务意义重大。目前已经有多种文法体系存在，其中影响最广泛、研究成果较为成熟的是最早由法国语言学家L.Tesniere在其著作《结构句法基础》(1959)提出的依存句法。他主张主要动词作为一个句子的中心支配其它

成分，而它本身不受任何其它成分支配。他主张主要动词作为一个句子依存句法的提出对语言学的发展产生了深远的影响，尤其是在计算机语言学界。Robinson J.J提出了依存关系的四大公理，奠定了依存句法的基础，这四条公理是：(1) 一个句子只有一个成分是独立的；(2) 其它成分直接依存于某一个成分；(3) 任何一个成分都不能依存于两个或两个以上的成分；(4) 如果A成分直接依存于B成分，而C成分在句子中位于A和B之间的话，那么，C或直接依存于A，或者直接依存于B，或者直接依存于A和B之间的某一成分[?]。以句子“聪明的小华想出了一种解法。”为例，进行依存句法分析，

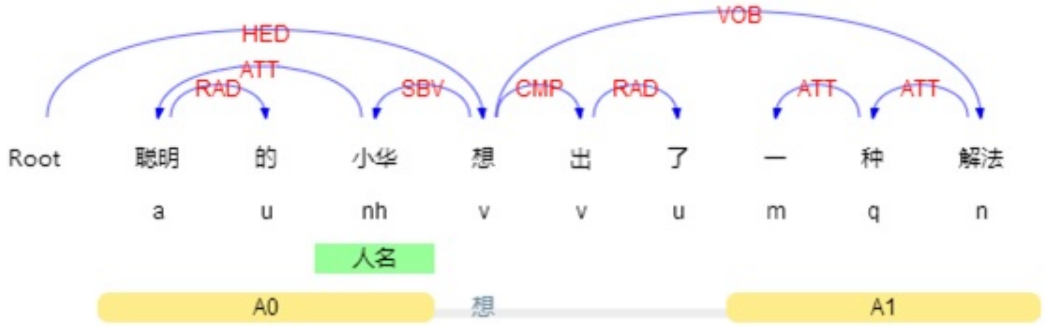

图 2.2: 依存句法树

整棵语法依存关系树以动词“想“为核心，按照依存关系逐步往下构建。目前汉语中的依存关系分类一般以清华大学计算机科学系的周明教授的分类方法为主，共有44 种依存关系[?]。包括SBV-主谓关系、VOB-动宾关系等。

2.5 RDF三元组

RDF全称Resource Description Framework，中文名资源描述框架，是一门基于XML的向万维网表达信息的语言，主要用于描述Web资源，使用RDF，人们可以使用自己的词汇表描述任何资源。

定义 2.16 RDF三个基本元素：

1. $URIs$: 用于指代资源

2. $Literals$: 数据值

3. $Blanknodes$: 空变量

定义 2.17 我们定义三个集合:

1. U：资源的集合

2. B：空变量的集合

3. L：数据值的集合

我们通过对它们的名字进行串联表示它们的并集，例如：UBL作为$UGBGL$的简写。RDF文件是一条条陈述的集合，每条陈述又叫做三元组，

定义 2.18 三元组的定义如下：

$$(s, p, o) \in UB \times U \times UBL \tag{2.1}$$

在这条triplet中，s是$subject$，p是$predicate$，o是$object$。

图Fig 3-2为其中一组RDF Triplet：

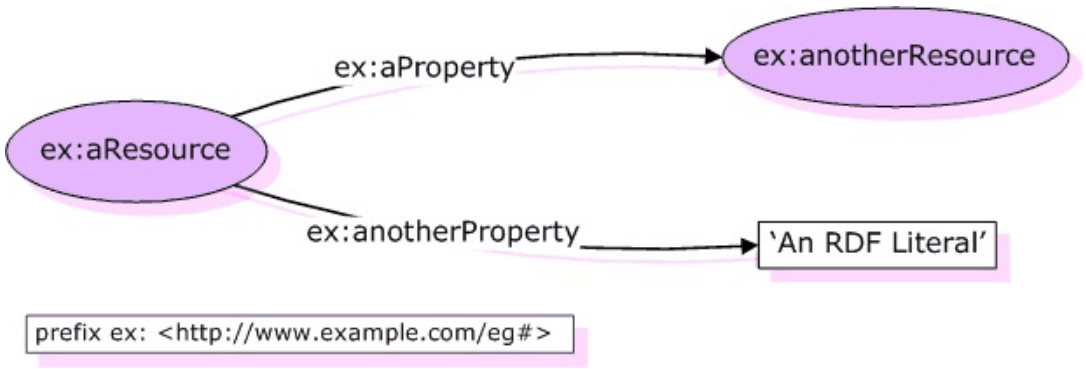

图 2.3: RDF triplet

2.6 TFIDF

TFIDF（term frequency－inverse document frequency）是一种用于资讯检索与文本挖掘的词频加权技术，属于统计学的一种方法，用以评估一个单词或字词短语对于一个文集或一个语料库中的其中一份文件的重要程度。字词的重要性

随着它在文件中出现的次数成正比增加，但同时会随着它在语料库中出现的频率成反比下降[?]。TFIDF的主要思想是：如果某个词或短语在一片文章中出现的频率TF高，并在其他文章中很少出现，则认为此词或者短语具有很好的类别区分能力，适合用来分类。[?][?]TF(Term Frequence) 意为词频，表达的是一个一个词在一个文档中出现的频率，IDF(Inverse Document Frequency) 意为逆向文件频率，TFIDF 认为如果一个词在一个文件中频繁出现，也就是说这个词的TF比较高，意味着这个词具有很好的代表性。同时这个词在整个文档集中出现的频率比较低，则表示这个词具有较好的类别区分能力。之所以使用词频而不是词数，是因为在较长的文件中，同一个词数出现的次数一般会更多。

命题 2.19

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \tag{2.2}$$

。

以上式子中$n_(i,j)$是该词在文件d_j 中的出现次数，而分母则是在文件d_j中所有字词数之和。

命题 2.20 IDF计算：

$$idf_i = \frac{|D|}{|j : t_i \in d_j|} \tag{2.3}$$

。

其中：

(1) $|D|$:语料中的文件总数。

(2) $|j : t_i Id_j|$:包含词语t_i的数目。

2.7 本体的关系

本体的体系结构包括三个要素：核心元素、元素之间的交互作用以及这些元素到语义规范之间的映射关系，Perez等对已有的本体进行分析，归纳出本体的基

本建模元语，其中最终要的两个是类和关系。类是相似术语所表达的概念的集合体，可以指任何事物，如工作描述、功能、行为、策略和推理过程等等。关系代表了在领域中概念之间的交互作用。本体中对概念之间的关系进行了区分，并把所有的关系分成四类，分别是part-of，kind-of，instance-of和attribute-of。

表 2.3: 本体关系分类

关系名	关系描述
part-of	部分与整体的关系
kind-of	概念之间的关系
instance-of	表达概念的实例和概念之间的关系
attribute-of	表达某个概念是另一个概念属性的关系

第 三 章　基于决断集模板进行溯因诊断

本体中的知识库以及规则是本体可以进行推理的前提。目前的本体构造主要由手工或者是半手工构建。由于本体的结构复杂，信息量大，因此本体的构建是一个长期的过程。在本体的构建过程中，构建的本体与观察值的不一致是一个常会发生的问题。因此，找到问题的原因，对观察值提出合理解释成为本体推理中的一个重要任务，这类任务也被称作是溯因推理问题。

3.1　解释与决定集

溯因推理的一个重要目的就是需要找出合理的解释对观察值的进行解释。一般的，这个解释不应该导致诊断本体的不一致，同时，结合本体中的背景知识，新构建出的本体能够蕴含观察值，且新构建的本体能够保持一致的特性。因此，我们对解释有以下定义：

定义 3.1 **(解释)** *给定一个一致的本体$\mathcal{O}$以及观察值α，$\mathcal{O} \nvDash \alpha$, 并且$\mathcal{O} \cup \{\alpha\}$ 是一致的，那么假如存在一个公理的集合$\mathcal{E}$使得$\mathcal{O} \cup \mathcal{E} \models \alpha$，$\mathcal{E} \nvDash \alpha$并且$\mathcal{O} \cup \mathcal{E}$ 是一致的，我们称这个集合$\mathcal{E}$为在本体$\mathcal{O}$中对α的解释。*

为了能够满足本体对蕴涵值的推理需求，解释在本体的溯因诊断中通常会以一种表达能力较高的DL语言比如$\mathcal{SROIQ}$。表达能力高的语言虽然可以满足本体对观察值的推理需求，但是较高的表达能力会带来另外一个问题，解释的空间会无限增大。为了能够尽可能提高解释的表达能力的同时能够限制解释的空间大小，本文利用模板来实现对解释的空间进行限制。在本体的推理中，决断集是一个重要的概念，对于一个一致的本体O以及观察值α，我们定义$Jst(\alpha, \mathcal{O})$为本体$\mathcal{O}$对观察值α的决断集的集合。因为决断集对蕴含值具有推理的合理性，同时决断集满足本体对观察值的解释需要是最小集合的约束，因此直观地本体对观察值的解释也会遵循相应的模板。

在定义决断集模板之前，我们首先需要可扩展公理，对于本体的一条公理，如果公理中的一个或多个的二元关系或者个体被二元关系变量或者个体变量替换，则这条公理会被称作可扩展公理。更多地，一条可扩展公理会被称为可全扩展公理如果这条公理的所有二元关系的个体都被二元关系中的个体变量替换。同时，可全扩展公理会保留本体中的$\top$，$\bot$以及个体的不变。在本体中，公理的类型种类多，由多条公理组成的决断集会产生数量难以接受的模板，因此我们需要限制模板的数量，也就意味着我们需要使用尽可能少的模板来表达尽可能多的决断集，同时，对于每个生成的模板，应该要有一种映射的方式使得决断集与之相对应。这种映射的方式我们称之为可扩展公理的替换，可扩展公理的替换能够把扩展公理的变量个体映射到个体名或者是个体变量。一个替换会被称作实例化替换如果这个替换能够把所有的个体变脸映射到个体名。

一般地，我们都需要这样的限制，对于每个由决断集产生的模板，都存在一个实例化的替换，使得这个由决断集产生的模板被映射到一个决断集中。但是这样的限制依旧会产生公理数不受限制的模板，考虑以下例子：

例 3.1 (解释) 令决断集$\mathcal{J}$是一直本体$\mathcal{O}$中对于二元关系断言$r_y(e_1, e_2)$的决断集，J_p是一个由决断集生成的决断集模板：

$$\mathcal{J} = \{r_M \circ r_N \sqsubseteq r_K, r_M(e_1, e_\mu), r_N(e_\mu, e_2)\}$$

$$\mathcal{J}_p = \{r_M \circ r_N \sqsubseteq r_K, r_M(e_{x_1}, e_{y_1}), \cdots, r_M(e_{x_n}, e_{y_n}), r_N(e_m, e_n)\}$$

从$\mathcal{J}$和$\mathcal{J}_p$可以看出，$\mathcal{J}_p$是一个公理基数没有上限的决断集模板，因为存在这样一个映射θ：

$$\theta = \{e_{x_i} \mapsto e_1, e_{y_i} \mapsto e_\mu, e_m \mapsto e_\mu, e_n \mapsto e_2 \mid 1 \leq i \leq n\}$$

由于这个θ满足条件所有的个体变量被映射到个体实例，因此这个映射θ是一个实例化替换使得$\mathcal{J}_p\theta = \mathcal{J}$。

为了避免出现决断集模板基数无限增大的情况，我们需要的决断集模板的替换做出进一步限制。在实例化替换的基础上，我们提出差异化实例替换。差异化实例替换不仅需要满足条件所有的个体变量都被映射到个体实例，还需要满足条件对于所有的不相同变量，被映射后的个体也不相同。直观地，一个差异化实例替换会是变量到个体间的一一映射。同时，一般地一个决断集模板需要能够生成至少一个决断集，然而为了保证最后生成的解释的合理性，我们需要决断集模板的所有差异化实例替换都只生成决断集，因此我们需要限制所有的对于$r(e_x, e_y)$决断集模板，在差异化实例替换的映射下能够维持蕴含$r(e_x\theta, e_y\theta)$这一特性。因此我们对决断集模板做出如下定义：

定义 3.2 (决定集模版) 一个由扩展公理组成的集合$\mathcal{J}_p$会被称为在本体$\mathcal{O}$中对于$r(e_x\, e_y)$的决定集模版如果$\mathcal{J}_p$满足以下两个条件:(1) 存在一个替换θ使得$\mathcal{J}_p\theta \in Jst\left(r\left(e_x\theta, e_y\theta\right), \mathcal{O}\right)$ (2)对于所有的差异化实例替换σ存在$\mathcal{J}_p\sigma \models r\left(e_x\sigma, e_y\sigma\right)$。

继续考虑例3.1中的决定集模板:

$$\mathcal{J}_p = \{r_M \circ r_N \sqsubseteq r_K, r_M(e_{x_1}, e_{y_1}), \cdots, r_M(e_{x_n}, e_{y_n}), r_N(e_m, e_n)\}$$

根据以上定义，在任一个一致的本体$\mathcal{O}$中，$\mathcal{J}_p$不是一个符合定义的对于二元关系断言实例$r_K(e_1, e_2)$的决断集模板。因为对于决定集模板$\mathcal{J}_p$的差异化实例替换θ,存在:

$$\{r_M \circ r_N \sqsubseteq r_K, r_M(e_{x_1}\theta, e_{y_1}\theta), r_N(e_m\theta, e_n\theta)\} \in Jst(r_M(e_{x_1}, e_{y_1}), \mathcal{O})$$

同时:

$$\{r_M \circ r_N \sqsubseteq r_K, r_M(e_{x_1}\theta, e_{y_1}\theta), r_N(e_m\theta, e_n\theta)\} \subset \mathcal{J}_p\theta$$

因此可以得出：

$$\mathcal{J}_p\theta \notin Jst(r_M(e_{x_1}\theta, e_{y_1}\theta), O)$$

同时，$\mathcal{J}_p$也不是$r_M(e_{x_1}, e_{y_k})$的决断集模板当$k > 1$。这是因为$\mathcal{J}_p\sigma \nvDash r_M(e_{x_1}\sigma, e_{y_k}\sigma)$，其中：

$$\sigma = \{e_{x_i} \mapsto e_1, e_{y_i} \mapsto e_2 \mid 1 \leq i \leq n, i \neq k\} \cup \{e_{x_k} \mapsto e_2, e_{y_k} \mapsto e_1\}$$

为了保证决断集能够被映射到决断集空间上，我们使用从决断集上生成决断集模板的方法来获取决断集模板。我们在生成的决断集的基础上，对二元关系中的个体进行变量替换，且对于不相同的个体名，我们使用不同的个体变量替换。我们对用不相同变量替换不相同个体的过程记作$lift(S, A, B)$。$lift(S, A, B)$表示对公理集合中的所有个体名，我们把A映射为变量X, B映射为变量Y, 其它不相同的个体分别映射到不同的个体变量。

命题 3.3 令$\mathcal{O}$为一致本体，$r_M(e_1, e_2)$是$\mathcal{O}$的一个蕴涵值，且$\mathcal{J}$是在$\mathcal{O}$中对于蕴含值$r_M(e_1, e_2)$的一个决断集。那么$lift(\mathcal{J}, e_1, e_2)$就是一个在$\mathcal{O}$中对于$r_M(e_1, e_2)$的一个决断集模板。

证明：(1)因为在决断集$\mathcal{J}$与$lift(\mathcal{J}, e_1, e_2)$之间具个体变量到个体实体的一对一映射，因此这里存在一个$lift(\mathcal{J}, e_1, e_2)$的差异化实例替换$\theta$使得$X\theta = e_1$，$Y\theta = e_2$，且$lift(\mathcal{J}, e_1, e_2) \cdot \theta = \mathcal{J}$。(2)令$\sigma$为$lift(\mathcal{J}, e_1, e_2)$的一个差异化实例替换，因此对于$lift(\mathcal{J}, e_1, e_2) \cdot \sigma$必然存在一个到$lift(\mathcal{J}, e_1, e_2) \cdot \theta$个体变量之间的映射$\rho$从而使得$X\theta = X\sigma$，$Y\theta = Y\sigma$且$lift(\mathcal{J}, e_1, e_2) \cdot \theta \cdot \rho = lift(\mathcal{J}, e_1, e_2) \cdot \sigma$。又因为$lift(\mathcal{J}, e_1, e_2) \cdot \theta \models r_M(X\theta, Y\theta)$，因此必然有$lift(\mathcal{J}, e_1, e_2) \cdot \theta \cdot \rho \models r_M(X\sigma, Y\sigma)$，考虑到$lift(\mathcal{J}, e_1, e_2) \cdot \theta \cdot \rho = lift(\mathcal{J}, e_1, e_2) \cdot \sigma$，因此可知对于所有的差异化实例替换$\sigma$都有$lift(\mathcal{J}, e_1, e_2) \cdot \sigma \models r_M(X\sigma, Y\sigma)$，命题成立。

要使得对于观察值α以及一致本体$\mathcal{O}$中的溯因解释$\mathcal{E}$遵循决断集模板，我们做出如下限制：$\mathcal{E}$是观察值α在$\mathcal{O} \cup \mathcal{E}$中的决断集的子集，其中这个决断集是由决断集模板通过差异化实例替换计算得来。我们把这个解释$\mathcal{E}$称作可接受解释：

定义 3.4 给定一个一致本体$\mathcal{O}$，一个可蕴含$r_M(X, Y)$决断集模板的集合$\mathcal{P}$以及观察值α，其中$\mathcal{O} \not\models \alpha$，解释$\mathcal{E}$如果满足存在一个决断集模板$\mathcal{J}_p \in \mathcal{P}$以及一个在决断集模板$\mathcal{J}_p$上的差异化实例替换$\theta$使得$\alpha = r_M(X\theta, Y\theta), \mathcal{E} \subseteq \mathcal{J}_p\theta$以及$\mathcal{J}_p\theta \in Jst(\alpha, \mathcal{O} \cup \mathcal{E})$，我们就把这样的解释称作可接受解释。

下面给出一个可接受解释例子：

例 3.2 考虑以下本体$\mathcal{O}$：

$$\{r_M \circ r_N \sqsubseteq r_K, r_M(e_1, e_3)\}$$

决断集模板$\mathcal{J}_p$为：

$$\{r_M \circ r_N \sqsubseteq r_K, r_M(A, e_w), r_N(e_w, B)\}$$

观察值α为：

$$r_K(e_1, e_2)$$

根绝决断集模板$\mathcal{J}_p$，我们可以得到$\mathcal{E} = r_M(e_1, e_3), r_N(e_3, e_2)$是一个对于观察值$r_K(e_1, e_2)$在$\mathcal{O}$中的可接受解释，因为存在一个差异化实例替换：

$$\theta = \{A \mapsto e_1, e_w \mapsto e_3, B \mapsto e_2\}$$

满足$r_o(e_1, e_2) = r_o(X\theta, Y\theta)$，$\mathcal{E} \subseteq \mathcal{J}_p\theta$且$J_p\theta \in Jst(r_o(e_1, e_2), \mathcal{O} \cup \mathcal{E})$。

给定一个一致本体$\mathcal{O}$以及一个观察α，其中$\mathcal{O} \not\models \alpha$，且$\mathcal{O} \cup \{\alpha\}$是一致的，那么本体的使用者更通常地会考虑以下三个特性：非平凡(i.e., $\mathcal{E} \not\models \alpha$)，一致(i.e. $\mathcal{O} \cup \mathcal{E}$是一致的)以及子集最小的(i.e., $\mathcal{O} \cup \not\models \alpha$ 对于所有的$\mathcal{E}' \subset \mathcal{E}$)。我们称满足以上三个条件的解释为合理解释。可以看出，在例子3.2中得到的解释$\mathcal{E} = \{r_M(e_1, e_3), r_N(e_3, e_2)\}$不是一个合理的解释。因为存在一个$\mathcal{E}$的子集$\mathcal{E}' = \{r_N(e_3, e_2)\}$)满足$r_o(e_1, e_2) = r_o(X\theta, Y\theta)$，$\mathcal{E}' \subseteq \mathcal{J}_p\theta$且$J_p\theta \in Jst(r_o(e_1, e_2), \mathcal{O} \cup \mathcal{E}')$。所以，我们结合以上条件定义如果一个可接受的解释满足非平凡，一致以及子集最小这三个条件，那么我们就称它为合理解释。

例 3.3 继续考虑例3.2，本体$\mathcal{O}$以及$\mathcal{O} \cup r_K(e_1, e_2)$也是一致的。正如前文提到的，$\mathcal{E}$并不是一个合理的解释，因为对于本体$\mathcal{O}$以及观察值$r_K(e_1, e_2)$，$\mathcal{E}$并不是一个最小集使得$r_o(e_1, e_2) = r_o(X\theta, Y\theta)$，$\mathcal{E} \subseteq \mathcal{J}_p\theta$且$J_p\theta \in Jst(r_o(e_1, e_2), \mathcal{O} \cup \mathcal{E})$。

相对的，$\mathcal{E}' = \{r_N(e_3, e_2\})$是根据模板集合$\mathcal{P}$在本体$\mathcal{O}$中对观察值$\alpha$的一个合理解释，使得$\mathcal{E}' \not\models r_K(e_1, e_2)$，$\mathcal{O} \cup \mathcal{E}'$是一致的且对于所有的$\mathcal{E}'' \subset \mathcal{E}'$不存在$\mathcal{O} \cup \mathcal{E}'' \models r_K(e_1, e_2)$。

在本体的诊断中，由于解释需要遵循决断集模板，而且不存在$\mathcal{J}_p\theta \subset \mathcal{O}$的情况，因此并非所有的变量个体都会被映射到本体$\{\mathcal{O} \cup \alpha\}$的个体集合中。在本文中，我们称这些变量为解释变量。由于解释变量的存在，如果想保证诊断的方法能够更加高效，一种考虑是只计算无法被其它解释通过实例化替换后得到的解释。然而解释变量的允许却会导致解释公理集合基数的变大甚至无法限制，考虑以下解释$r_o(A, A)$，这个解释可以由解释$\mathcal{E} = \{r_o(A, e_{x_1}), r_o(e_{x_2}, e_{x_3}), \cdots, r_o(e_{x_{n-1}}, e_n)\}$通过映射$\theta = \{e_{x_i} \mapsto A \mid 1 \leq i \leq n\}$获得。因此对于解释的替换，我们提出一个新的概念，叫做解释差异化替换。解释$\mathcal{E}$的差异化替换会可以把$\mathcal{E}$的解释变量映射到不同的解释变量或者是一个没有在$\mathcal{E}$中已经出现的变量。在前文提到的例子中，θ并不是一个解释的差异化替换，因为e_{x_i}被映射到的变量A是解释$\mathcal{E}$中已经存在的个体实例，所以θ不是解释的差异化替换。因此我们提出一个新的子集概念：

定义 3.5 ($\subseteq_{\mathsf{ds}}$-*minimal* **解释**) 一个解释$\mathcal{E}$会被称为$\subseteq_{\mathsf{ds}}$-*minimal* 如果这个$\mathcal{E}$满足以下条件：不存在这样一个解释$\mathcal{E}'$使得$\mathcal{E}' \subseteq_{\mathsf{ds}} \mathcal{E}$ 且$\mathcal{E} \not\subseteq_{\mathsf{ds}} \mathcal{E}'$，其中$\mathcal{E}' \subseteq_{\mathsf{ds}} \mathcal{E}$ 表示存在一个$\mathcal{E}'$的差异化替换θ，使得$\mathcal{E}'\theta \subseteq \mathcal{E}$。

$\subseteq_{\mathsf{ds}}$-*minimal*解释是一个比子集最小更强的概念，以下定理说明$\subseteq_{\mathsf{ds}}$-*minimal*是子集最小的充分条件：

命题 3.6 一个对于观察值α在本体$\mathcal{O}$中的$\subseteq_{\mathsf{ds}}$-*minimal* 解释$\mathcal{E}$同时也是一个*subset-minimal*的解释。

证明：对于所有的$\mathcal{E}' \subset \mathcal{E}$，我们有$\mathcal{E}' \subseteq_{\mathsf{ds}} \mathcal{E}$且$\mathcal{E}' \not\subseteq_{\mathsf{ds}} \mathcal{E}$。所以，如果不存在对于观察值$\alpha$在本体$\mathcal{O}$中的解释$\mathcal{E}'$使得$\mathcal{E}' \subseteq_{\mathsf{ds}} \mathcal{E}$且$\mathcal{E}' \not\subseteq_{\mathsf{ds}} \mathcal{E}$，那么也不会存在$\mathcal{E}' \subset \mathcal{E}$使得$\mathcal{O} \cup \mathcal{E}' \models \alpha$，所以命题成立。

验证解释$\mathcal{E}$是否是$\subseteq_{\mathsf{ds}}$-*minimal*不需要考虑解释$\mathcal{E}$的所有子集组合排列的情况。实际上我们只需要考虑n个更小的解释即可。这里的n等于$\mathcal{E}$中公理数的基数以及个体实例的总和。在介绍这两种方法之前，我们需要先介绍两种集合概念。一种是公理集S的最近变量集$lifts_1(S)$，最近变量集$lifts_1(S)$使用变量替换一个在S中已经存在的个体实例。例如：对于$S = \{r_o(A, B)\}$存在两个最近变量集分别是$\{r_o(e_x, B)\}$和$\{A, r_o(e_x)\}$。然后，我们定义一个最大真子集，最大真子集有且仅有S中的某一公理外的所有公理。我们把最大真子集记作$subs_1(S)$。以下命题展示了验证解释$\mathcal{E}$是否是一个$\subseteq_{\mathsf{ds}}$-*minimal*解释的方法。

命题 3.7 *对于观察值α在本体$\mathcal{O}$中的解释$\mathcal{E}$是一个$\subseteq_{\mathsf{ds}}$-minimal解释当且仅当$\mathcal{E}$满足以下条件：$\mathcal{O} \cup \mathcal{E}' \not\models \alpha$对于所有的$\mathcal{E}' \in subs_1(\mathcal{E}) \cup lifts_1(\mathcal{E})$。*

3.2 RDFS构建框架

通过对本体学习以及RDFS相关的知识的研究，综合考虑到中文自然语言的特点、RDFS的表达方式以及本体的基本元素需要，设计出了一个基本的从中文自然语言文本构建基于RDFS本体模型的系统框架。

3.2.1 目标本体特性

要从中文自然语言文本中构建出准确的RDFS本体模型，该系统框架应该具有且不限于以下几点特征:

(1) 能够以较高的精度完成中文的自然语言处理

(2) 能从信息源中识别出领域概念

(3) 能自动抽取出分类层次关系

(4) 能抽取出概念之间的非分类关系

基于以上几点，可以设计出系统主要流程为：读取输入信息源－－自然语言处理－－本体概念识别－－概念以及关系抽取－－元素转换－－构建RDFS模型。

3.2.2 中文文本预处理

本文设计的系统框架的输入信息源为中文自然语言文本，为了进行术语识别

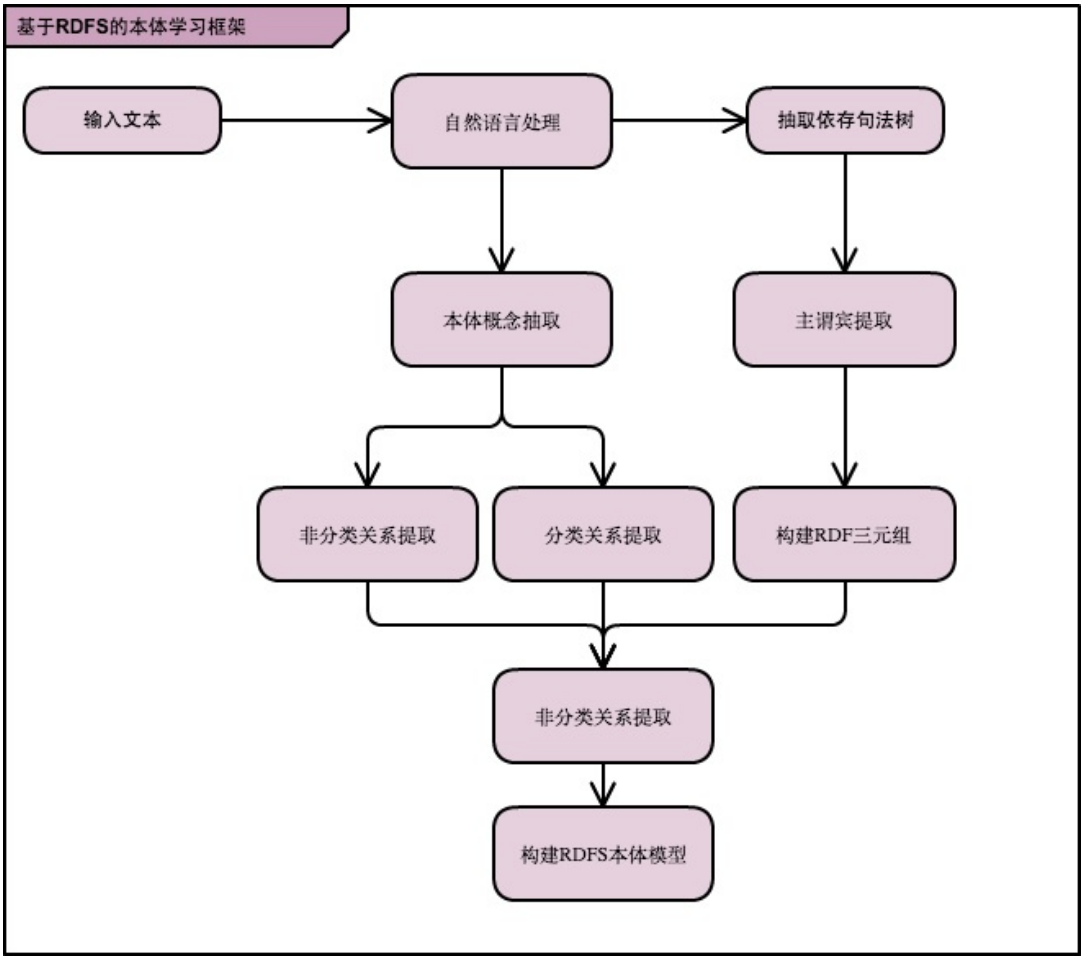

图 3.1: 基于RDFS的本体学习框架

以及该概念关系抽取的相关操作，需要对输入信息进行预处理，包括分词、语义角色标注、依存句法树生成等工作。目前国外的自然语言处理框架在完成自然语言处理的基本任务时能有较高的准确率，但是这些框架要么不支持中文，要么在中文环境中处理的准确率较低，影响系统的准确性。在国内目前也已经研发出面向中文的自然语言处理框架，其中已经发展得比较成熟的有哈尔滨工业大学的nlp和复旦大学的nlp，这两者对中文自然语言的处理都有较高的准确率，鉴于哈尔滨工业大学nlp需要通过WebAPI的方式进行调用，而复旦大学的fnlp可以在本地以第三库的形式进行调用，在使用大规模文本作为信息输入源进行测试的时候会有相对较好的速度性能表现。所以本文选择使用由国内复旦大学研发的中文自然语言框架fnlp来进行中文文本的预处理工作。

3.2.3 本体概念的识别

术语是一种结合紧密的固定或半固定的词或短语，具有结合紧密型和语言完备性特点，进而，它还是一种具有很强的领域特征的词语，具有领域性[?]。在本体中，术语与概念在概念上是一致的，概念的自动提取是构建本体的重要工作之一[?]。在目前的本体构建中，领域概念的抽取主要有两类方法，分别是基于统计的方法和基于语义处理的方法。语义处理通过对文档进行预处理，如分词、语义标注、语法树分析、依存结构生成、实体识别等，对信息源进行结构与模式分析，提取出术语。通过语义分析抽取术语不需要大量的信息源，相对较少的计算量，但是语义分析的方法依赖于严格的语法结构，对于口语化的表达难以达到理想的效果，对输入信息源的质量有着较高的要求。而基于统计学的方法并没有这些限制，基于统计学的方法不需要输入的信息源严格遵循语法结构，能够比较容易地提取出未登录词。与之相对的，统计学的方法也有缺点。统计学的方法在抽取低频术语的时候容易遗漏，且统计学需要大量的输入信息源。推荐度的计算主要基于统计学方法中的词频逆文档频率(term frequency – inverse document frequency)。TF.IDF用于对词语的区分能力进行计算，分数越高，意味着词语的区分能力越强，表达范围归入该领域的可能性就越大[?]。

命题 3.8 词频逆文件频率的计算:

$$tf_{i,j}idf_i = \frac{n_{i,j}}{\sum_k n_{k,j}} \frac{|D|}{|j : t_i \in d_j|} \tag{3.1}$$

。

词频逆文件频率对单位字词有较好的计算能力，但是在计算时难以对复合型短语进行计算。以“书画艺术”为例，经过自然语言处理之后，会被分词模块识别成“书画”和“艺术”两个词语，而“书画艺术”是在艺术领域里一个重要的术语。在Dennis, B等人对术语的研究中，他们证明了术语在主要由名词和名词短语表达，在名词短语中，又以双词组合占多数[?]。

定义 3.9 以一个或者两个字词组成的有完整意义的短语，我们称为base-item。

在中文中，通过对语料的文档人工处理，我们提取出三组典型的base-item的语义形式，[名词]，[形容词，名词]，[名词，名词][?]。根据这三组语义形式，我们可以过滤出候选的术语集，包括字词和短语，然后再使用统计学方法进行筛选，具体步骤如下：

- 对文档按照标点符号进行分句操作，对于非规则或者无符号文本部分，利用空格与换行等符号进行断句。
- 利用fnlp对句子进行分词操作，并为每个字词标注词性。
- 通过应用语义表达规则从字词集中提取出候选的术语。
- 计算出候选术语的推荐度，制定阀值，筛选出最终的术语集。

3.2.4 分类关系的抽取

概念之间的关系是本体构建中的重点。本体通过关系把各种概念连成一体，以此表达领域之中概念之间的联系。在本体中的关系主要有两类，分别是分类关系和非分类关系。这两类关系在文本中不管是体现的方法还是形式都不一样，因此对于这两种关系的提取需要使用不同的方法分别进行操作。系的方法主要是基于Hearst模式的抽取方法。Hearst模式是一种基于模式的信息抽取方法[?][?]，它通过识别句子中的特定模式来抽取概念之间的关系。由于自然语言句式较多，表达方式多样，因此准确率较低。而且Hearst 模式是基于西文的，在中文上并不适用。中文的语法相较英文要复杂不少，从模式上进行信息抽取的难度较大，因此本文通过结合Wordnet的方法进行信息抽取。Wordnet中保存了与词原有上下位关系的词，通过查询中文Wordnet可以对领域概念之间的上下位关系进行验证。

3.2.5 非分类关系的抽取

非分类关系与分类关系不同，非分类关系可以相对容易的根据文本中进行提取。GOLF系统中对非分类关系的抽取采用的是关联规则的VCC(n)事

务方法，VCC(n)事务方法假定：如果概念c1和c2间具有非分类关系v，当且仅当c1和c2 都出现在带有动词v的n个词内(即c1和c2都出现在动词v的周围)，可以用一个条件概率来表示动词和概念对之间的关联度。这种方法虽然可以从文本中提取出大部分的关系，但是缺点是会抽取出太多无意义的和错误的关系，这种情况在长难句中更加明显。考虑以下例子：戴着帽子的老师在看书。(1) 这是一个比较常见的句式，在VCC(n) 的提取方法中由于“帽子”与“书”都在动词“看”的附近，因此“帽子”与“书”的关系得分会比较高，而这种关系并不是我们需要的。针对以上情况，本文选择利用语义分析的方法来提高关系抽取的准确度。在汉语中，谓语是表达两个概念之间关系的最直接和最主要的方式，而“主谓宾”则是汉语中最基本的句式，其中“主宾”相当于一个关系实例中的两个概念，主、谓、宾模式构成的三元组可以为领域关系的抽取提供强有力的参考[?]。主谓宾的模式(subject, Verb, Object)类似于RDF 模型中的(Subject, Predicate, Object) 三元组，本文会使用RDF三元组的形式进行表达。主谓宾的提取首先需要借助fnlp对句子进行自然语言预处理，生成依存句法树，依旧以句子(1)为例，生成该句子的语法生成树： 从依存句法中提取主谓宾的算法如下：输

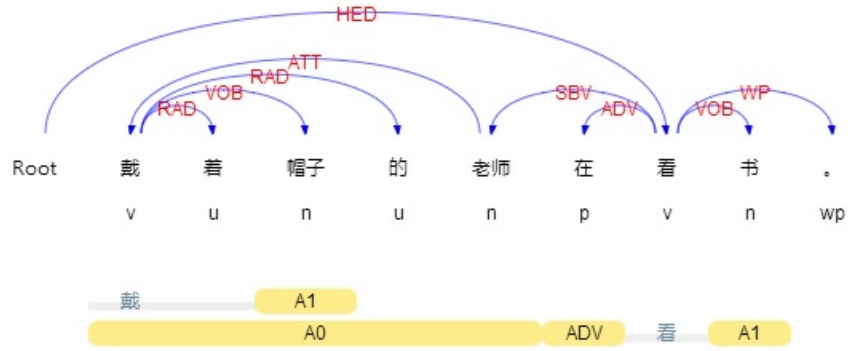

图 3.2: 基于RDFS的本体学习框架

入：中文句子s。输出：句子主谓宾结构1. 遍历句子中的元素，找到HED词。2. 提取出依存HED词的词集D。3. 遍历D，找到与HED有SBV关系的词S，若无，返回空。4. 继续遍历，找到与HED有VOB关系的词O，若无，返回空。5. 返回(S,

HED, O)组成的主谓宾三元组。

Input: Dependency Tree t
Output: $svotriplet(r)$
for *each* $word$ *in tree* t **do**
  res instance of svotriplet;
  if $word$ *is HED* **then**
    $wordset$ = words depend on $word$;
    $res.HED = word$;
    for *each* w *in* $wordset$ **do**
      if *relation between* w *and* $word$ *is* SBV **then**
        $res.subject = w$;
      end
      if *relation between* w *and* $word$ *is* VOB **then**
        $res.object = w$;
      end
    end
  end
end
return res;

Algorithm 1: SVO Extraction

对Fig 3.2中的句法树进行操作，分别找到与HED词“看”有SBV依存关系的“老师”与有VOB关系的“书”，可以提取出句子(1)的主谓宾结构三元组(老师, 看，书)，而这就是我们需要的一条RDF关系实例。

3.2.6 基于RDFS的本体模型的构建

本体中的关系主要有是part-of，kind-of，instance-of和attribute-of[?][?]。这四个关系在RDFS中并没有具体指明两者之间的关系，但是我们可以在RDFS中找到表达方式与之相近的关系。在RDFS中，rdfs:type表示的是类与实例之间的关系，我们用来表示本体中instance-of的关系，同理，被用来表示概念之间的继承关系会映射到被用于表达父类与子类关系的rdfs:subClassos，而attribute-of则会使用RDF三元组中的property代替。至于part-of，由于RDFS中并无与之相近的表达，且part-of关系可以看做是attribute-of关系的细化，所以我们把part-of关系也

映射到RDF三元组中的property来对其进行表达。最后的关系映射关系如图Fig 3.3：

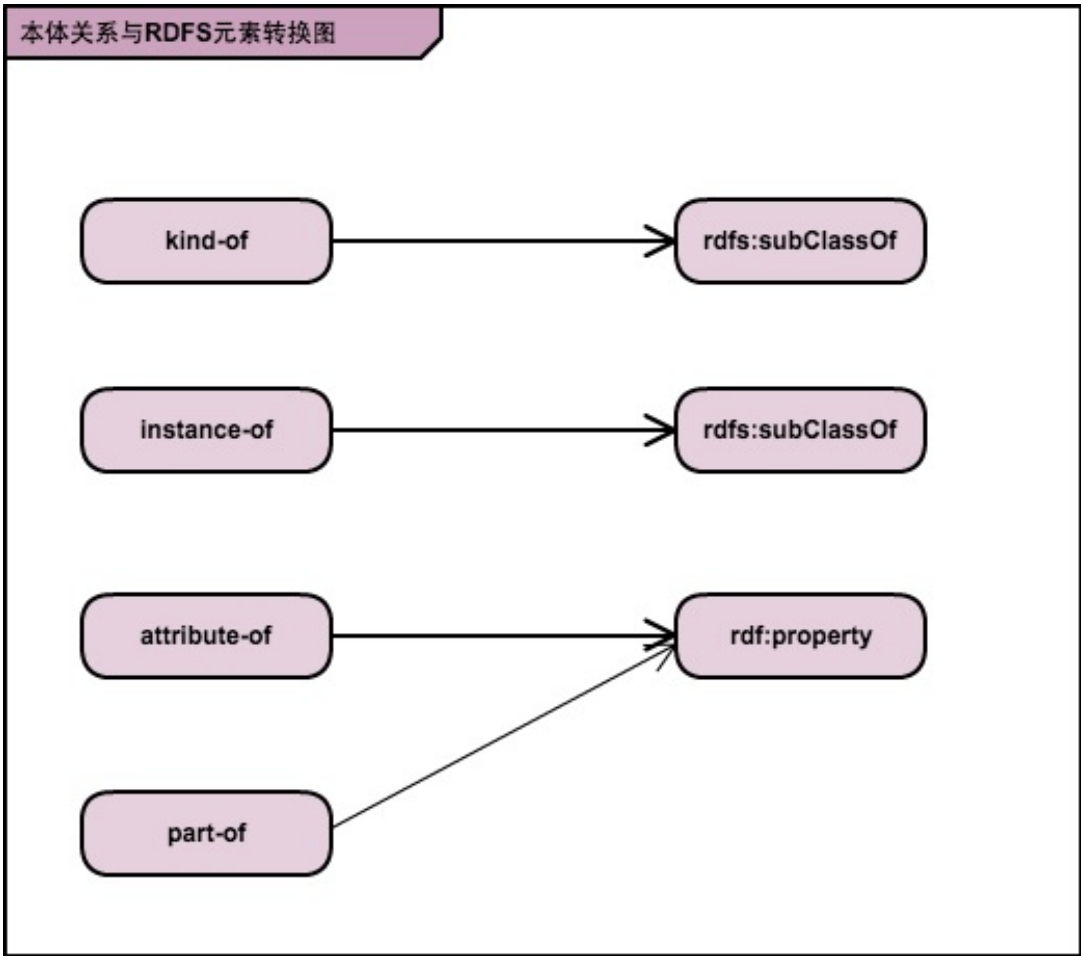

图 3.3: 基于RDFS的本体学习框架

第 四 章　实验与分析

我们在本文介绍了一种从中文自然语言文本构建RDFS本体的框架，其中包含了术语识别与关系提取中用到的多种方法。在这一章当中，我们设计了一个系统Text2RDFS(以下简称t2r)对上文中的理论进行检验，并对比了不同方案表现的差异性。

4.1　实验基本情况

本实验系统t2r可以从中文语料库中构建出RDFS本体，实验系统使用java编写，自然语言预处理方面使用的是复旦大学自然语言处理库fnlp。 t2r主要分为

表 4.1: 实验环境

环境	描述
CPU	1.4 GHz Intel Core i5
RAM	4GB
OS	OS X Yosemite 10.10.3
自然语言处理库	复旦自然语言处理库fnlp

两个模块，第一个模块是术语提取模块，在t2r系统中我们分进行了使用语义分析法和使用统计法提取术语的实验。第二个模块是关系提取模块，在t2r系统中我们实现了非分类关系和分类关系的抽取模块。分类关系的提取采用的是借助wordnet进行上下位关系查询的方法，非分类关系的提取则是对本文第三部分介绍的方法进行实践。对于t2r的语料库，我们在进行了仔细对比分析之后，选择了由复旦大学李荣提供的文本分类语料库中的测试预料。该语料库由人工分类，总数9833篇，共20个类别。其中的语料文章大多数来自于国家官方媒体的新闻通稿或者学术期刊中的文章，因此，这些文章具有较高的语法正确率，避免了口语化、网络化表达对实验测试的影响。文章中有合适比例的长句和短句，可检验t2r系统在不同环境中的性能表现。同时，语料库中的文章具有较高的时效性，

一定程度上避免了构建出的RDFS模型过时的情况。在领域的选择上，我们挑选艺术领域的语料库进行测试。这三个领域具有广泛性，语料数量多，避免了个别极端情况对实验的影响。

4.2 实验结果与分析

由于t2r的设计中上下位关系的识别主要依靠现有的中文wordnet，因此对t2r的评价集中在领域术语识别和非分类关系提取两个模块的表现上。对于领域术语识别，我们使用IE领域广泛使用的准确率(precision)、召回率(recall)[?]对t2r术语识别模块的识别结果进行评价。我们这次总共测试了43篇文档，其中共有人工识别出的术语1236个。

表 4.2: 统计方法的领域术语提取结果

阀值(%)	人工分析正确数	召回率(%)	准确率(%)
1	127	10.8	38.2
2	221	17.9	33.0
3	301	24.4	30.0
5	351	28.4	21.0
10	546	44.2	16.3

表 4.3: 结合语义分析法统计方法的领域术语提取结果

阀值(%)	人工分析正确数	召回率(%)	准确率(%)
1	185	15.0	55.4
2	381	30.8	57.0
3	512	41.4	51.1
5	638	51.6	38.2
10	773	62.5	23.1

从表4.2和表4.3可以看出，简单的统计学的方法难以在准确率和召回率之间取得较好的平衡，在保证足够回收率的同时难以保持较高的准确率，若要保持足够

高的准确率，就不得不以较低的回收率为代价。而在结合了语义分析的相关方法之后，在相同阀值的情况下，准确率和回收率都获得了一定程度上的提高。

在非分类关系抽取方面，我们难以统计所有的关系数量，因此我们以准确率来对结果进行评价。 关系抽取中我们基于主谓宾的方法在处理短句集方面取得

表 4.4: 基于统计方法的领域术语提取结果

文档类型	文档数量	关系抽取数量	准确数	准确率(%)
长句集	20	397	246	62.0
短句集	32	683	584	85.5

了较好的效果，得到了较高的准确率，而面对长句的时候，由于中文长句的关系复杂性，准确率有所降低。在整体来看，我们的关系抽取模块有着相对优秀的表现。

第 五 章　总结与展望

5.1　工作总结

现今阶段，语义网的高速发展以及它的潜在价值获得了越来越多人的关注，作为语义网的重要组成部分，本体的构建已经成了许多研究的重点。在国外，本体构建的研究已经取得了不少的进展，例如Delia Rusu 等人提出的基于Treebank Prser的Triplet提取算法可以比较准确的从语句中提取术语之间的关系了。遗憾的是，中英文之间的差别影响了这些学术成果的可复用性，因此，国内的相关研究人员开始对本体的自动和半自动构建进行研究，但是目前进展缓慢。基于现有的研究成果，本文从中文语料库中进行领域术语识别、分类关系和非分类关系提取，实现了从中文语料库中初步构建出以领域本体的RDFS模型。同时我们还分别比较了利用统计学方法和语义分析法对领域术语进行识别，经过试验寻找出不同方法的使用场景。在进行关系抽取的时候，我们利用代词和零代词的消解提取出隐藏的关系，提高了抽取的准确率和覆盖率。

5.2　不足与改进方向

在本文的本体构建中，由于中文短语组合的复杂性，尚不能对以短语形式出现的领域本体术语进行比较准确的提取。术语的缺少会导致关系抽取数量的减少，影响了构建出的本体的全面性。在未来的研究中，主要由两个方面需要进行改进。一个是对上下位关系的识别，目前的方法依赖于中文wordnet等语义词典的帮助，而目前中文语义词典的覆盖尚不全面，遗漏关系的现象比较严重。同时，在本次实验的基础上，通过对语言模型的应用抽取出本体中规则的推倒，将使构建出的本体的可用性提高到一个新的层次。

致　谢

不知不觉就走完了大学本科四年的历程。在四年的大学生活中，我遇到了很多优秀的老师，他们在传授知识上尽心尽力，学士渊博的他们也是我学习的榜样，他们治学严谨，对知识渴求，对学生负责，让我走进知识的殿堂，对此我要对他们表示衷心的感谢。要特别感谢万海老师的悉心指导，从大一开始上万老师的课，万老师对我的严格要求为今后的学业道路打下坚实的基础。在保研之后，老师就开始对我进行了科研方面的相关指导，带领我提早开始进行研究生的工作，多次在我存有疑惑的时候耐心的替我解决问题，并且总能够在最及时的时候给予回复，在此要特别感谢。

要感谢我的舍友们，在撰写本文时多次进入熬夜状态，晚上的灯光和键盘声多少对他们的休息产生了影响，感谢他们给予的理解。

还要感谢我的爸爸妈妈，多年的养育之恩终于让我长大成人，他们在我低谷的时候给予我关怀与帮助，在我开心的时候一起分享快乐，这份来自家庭的感动一直是我前进的最大动力。

最后要感谢和我一起度过这四年的朋友和同学，因为你们让我这四年更加的珍贵。

参考文献

附　录

中英文词汇对照表

Resource Description Framework(RDF)	资源描述框架
Resource Description Framework Schema(RDFS)	资源描述框架模式
DARPA Agent Markup Language(DAML)	DARPA代理标记语言
Ontology Inferrence Layer(OIL)	本体建模语言
Web Ontology Language(OWL)	网络本体语言
Extensible Markup Language(XML)	可扩展标记语言

原创性声明

本人郑重声明：所呈交的学位论文，是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已经注明引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的作品成果。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。本人完全意识到本声明的法律结果由本人承担。