Simple Ontologies for Practical Information Extraction and Advanced Information Extraction for Practical Ontologies

简单本体在实用信息抽取中的使用 及 针对实用本体的高级信息抽取

Hong Li （李宏）
DFKI, LT-Lab（德国人工智能研究中心语言技术实验室）
Alt-Moabit 91c
10559 Berlin
Phone: +49-30-3949-1836
Email: lihong@dfki.de

Yi Zhang（章毅）
DFKI, LT-Lab（德国人工智能研究中心语言技术实验室）
Stuhlsatzenhausweg 3
66123 Saarbruecken
Phone: +49-681-85775-52xx
Email: yizhang@dfki.de

Feiyu Xu（徐飞玉）
DFKI, LT-Lab（德国人工智能研究中心语言技术实验室）
Alt-Moabit 91c
10559 Berlin
Phone: +49-30-3949-1812
Email: feiyu@dfki.de

Hans Uszkoreit (汉斯．乌兹库赖特)
DFKI, LT-Lab（德国人工智能研究中心语言技术实验室）
Stuhlsatzenhausweg 3
66123 Saarbruecken
Phone: +49-681-85775-5282
Email: uszkoreit@dfki.de

Abstract

Information extraction can be regarded as a pragmatic approach to semantic understanding of natural language texts. Ontology is very important for modeling and specifying knowledge e.g. relations between entities and concepts. Therefore, ontology is often used for definition of the information extraction tasks. The advanced information extraction technologies such as complex relation extraction can be utilized for learning language patterns, which can recognize ontological relations from the free texts and extract relation instances. In this paper, we will describe an ontological model for information extraction tasks and present a general machine learning framework DARE for learning relation extraction patterns and extracting relation instances. DARE system has been intensively used for the English language. In this paper, we apply DARE to the Chinese newspaper texts and detect Chinese relation extraction rules and relation instances. Furthermore, we will compare our experiments with Chinese texts with the English texts.

信息抽取是用来理解自然语言文本语义的一种非常实用的方法。而本体（ontology）则对知识（例如实体和概念间的关系）的定义与建模起重要作用。因此本体经常被用来定义信息抽取任务。 高级信息抽取技术，例如复杂的多元关系抽取，可用于学习语言模板，进而从自然文本中识别出本体定义的关系， 并抽取出这些关系的实例。本文将描述一个为信息抽取任务而设计的本体模型，以及一个关系抽取的框架系统 － DARE。DARE 使用了机器学习方法， 可以自动学习关系抽取的语言模板，并应用这些模板来抽取关系实例。DARE 系统已被深度应用于英语文本的关系抽取。本文将进一步使用 DARE 来处理中文新闻文本， 从中学习语言模板和抽取关系实例，并与英文处理的结果进行比较。

Keywords: ontology, information extraction, relation extraction, automatic rule learning

本体， 信息抽取，关系抽取，规则自动学习

1 概述

本体（Ontology）在信息技术中起着组织与分类知识和信息的关键作用。在本体的各类应用中，最具潜在价值的是语义网络（Semantic Web)，一个利于语义访问的结构化万维网（Berners-Lee 1999)。另一方面， 信息抽取（Information Extraction）技术旨在从自然语言文本中抽取结构化的信息 （Grishman and Sundheim 1996；Appelt and Israel 1999）。抽取的信息通常包括人名、地名、组织机构等概念实体及其之间的关系，如亲属关系、雇用关系，等等。这些实体及其之间的关系，正是本体的基本组成部分。因此本体常用于定义信息抽取系统的任务目标，组织和描述抽取出的信息。与此同时，信息抽取也被应用于更新、填充和扩展已有的本体知识库（Hearst 1992）。

历经二十世纪九十年代的多次信息理解会议MUC[1]（Grishman and Sundheim 1996；Appelt and Israel 1999）及其后的自动内容抽取项目ACE[2]，信息抽取已经成为一个被广泛认可的研究领域。在过去的十年里， 信息抽取技术已由单领域的复杂事件抽取转向标准的跨领域基本实体识别、基础语义关系与事件抽取。 MUC和ACE定义的实体和关系只基于一个简单的语义模型， 并对应一个简单的本体模型（Appelt 2003）。Xu（2007）提出了一个扩展的语义模型，用于构建简单的本体和实用的信息抽取任务。以此为基础，Krause等（2012）进一步形式化地定义了二元及多元的实体关系。

领域适应（domain adaptation）是信息抽取中的核心研究方向之一。一个被普遍应用的方法是自动学习用于识别目标关系实例的分类模型或语言模板（Agichtein and Gravano 2000； Yangarber，et al. 2000；Ravichandran and Hovy 2002；Stevenson and Greenwood 2005）。一个更高效的针对领域适应的策略是使用一种被称为“Bootstrapping”的最小化监督式机器学习(minimally supervised machine learning)方法，即从非常有限的初始知识开始进行多次的机器自动迭代学习。这种方法无需手工标注学习语料或构建大规模的初始知识库。它从有限的初始关系实例或关系抽取模版出发，自动迭代标注文本，逐步发掘新的关系实例和语言模版。本文将介绍一个称为DARE[3]（Xu 2007）的信息抽取系统。DARE系统实现了这种最小化监督式机器学习方法。与其它系统相比，DARE能够应对相对复杂的关系，并自动学习提示语义关系的语言模版、为句子成分分配关系中的语义角色。Uszkoreit（2011）以DARE为例， 系统地分析了最小化监督式机器学习中的参数对信息抽取系统性能的影响。目前DARE已被用于多个领域中的信息抽取，如获奖事件、社会人际关系、管理交替事件，等等（Xu 2007；Xu，et al. 2007；Xu，et al. 2010；Krause，et al. 2012）。以往的DARE实验主要是在英文的新闻类和网络文本上进行的。本文将首次尝试在中文新闻类文本上运行DARE系统。

本文的结构如下：在第2节中，我们将引入一个语义模型，并基于该模型描述一个用于实用信息抽取的简单本体。第3节将主要介绍了DARE系统的体系框架，及其在英语文本上的关系抽取。在第4节， 我们将描述DARE在中文新闻类文本上抽取目标关系的几个实验，并在第5节介绍和分析实验评测结果。第6节总结并提出未来研究方向。

---

[1] 英语：Message Understanding Conferences，简写为 MUC

[2] 英语：Automatic Content Extraction program ，简写为 ACE

[3] 英语：Domain Adaptive Relation Extraction，可适应领域关系抽取

2 简单本体在实用信息抽取中的使用

2.1 语义模型

信息抽取可被视为能够理解自然语言语义的一种非常实用的方法。多次的MUC会议及其之后的ACE已经初步确定了信息抽取任务的语义模型，定义了其中的一些相关组件（Appelt 2003）。Xu（2007）进一步扩展了这个语义模型，提出这个模型应包含如下组件：

* 实体：在文本中提到的存在于现实世界中的个体
    - 简单实体：单个对象
    - 集体实体：文本中明确提到的同类型对象的集合
* 关系：在多个实体之间的相关连的特性
* 复杂关系：实体与关系之间的关系
* 属性：描述个体特性的一元关系
* 时间点或时间段：关系可不受时间限定，也可以与特定的时间点或时间段绑定
* 事件：特定的一类涉及至少一个关系变化的实体间的简单或复杂关系

扩展后的模型定义了复杂的关系，并强调了时间属性和关系之间的联系。例如，人的性别（通常）属于不受时间限定的属性，而人的年龄则是与时间点绑定的属性。父子关系是不受时间限定的二元关系，而雇佣关系则与时间段绑定的二元关系。

我们从语义角度出发，以本体为基础归纳总结了信息抽取的实际任务：

* 针对应用领域建立本体模型
    - 定义相关的领域概念及其之间的IS-A和PART-OF关系
    - 定义相关的领域概念之间的领域特有关系
    - 定义相关的领域事件
* 识别语言学实体
* 对语言学实体进行分类并对应到相应本体中的语义类别
* 识别等价语言学实体
* 识别语言学结构
* 对语言学结构的语义进行分类
* 对语言学实体在关系和事件中的语义角色进行分类

2.2 针对诺贝尔奖获奖关系的简单本体

我们以诺贝尔奖获奖关系为例构建一个本体模型。 建模方法详情可参阅（Frank，et al. 2006；Xu，et al. 2006）。

一般而言， 开始建立应用领域的本体模型时可以参考现有的通用本体。除了ACE的本体模型及类似资源外，以下两种资源可以作为候选参考：1) 基于知识工程（Knowledge Engineering）的上层本体模型SUMO（Pease and Li 2002）及其中层规格说明MILO（Niles and Terry 2004）；2) 结构化的词汇库WordNet（Miller，et al. 1993）。Niles和Pease（2003）构建了SUMO中的抽象概念与WordNet词义间的对应关系。我们将以SUMO的本体模型作为基本主干，利用上述SUMO与WordNet词义之间的对应关系来定义针对获奖关系的本体中的一些重要的子概念。

我们应用领域中的主要概念有获奖者、奖项、 奖项分类及一些非领域特有的概念与实体，如时间／日期、金额、组织、人、地点等。表2.1中罗列了一些获奖领域特有的概念与SUMO概念之间的对应关系。例如，获奖者对应SUMO中的cognitiveAgent，主要被它的两个子概念human及organization所继承。而奖项分类的子概念（“和平”奖项除外）同时也是SUMO中fieldOfStudy的子概念，如“化学”奖项。每一个概念进一步拥有其特有属性，如“人”被赋予“姓”“名”等属性。这些概念按照层次化的关系被组织在一起。

表2.1: 获奖领域中的概念与SUMO概念之间的对应关系

概念类型	获奖领域	SUMO
实体	奖项	award，…
实体	获奖者	cognitiveAgent
实体	人	human
实体	组织	group
实体	奖项分类	fieldOfStudy，…
事件	诺贝尔奖获奖	unilateralGetting
事件	诺贝尔奖提名	declaring, deciding

与实体的分类类似，关系和事件可以通过IS-A关系组织起来（Xu，et al. 2006）。例如，获奖关系中的一般事件总类称为receive-award。它包含了如下的论元：获奖者、获奖内容、获奖原因、地点和时间。该事件的确切定义如下：

获奖者(recipients)	人员或组织的集合
获奖内容 (award)	奖项、奖金、奖牌、称号…
获奖原因 (reason)	成果（成就，能力，发明……）
获奖地点 (location)	地区
获奖时间 (time)	日期

receive-prize事件是receive-award事件的一个子类，因此包含了一些更具体的论元和定义：

获奖者(recipients)	人员集合
获奖内容 (award)	奖项
获奖原因 (reason)	成果（成就、能力、发明……）
奖项分类 (area)	奖项领域、类别
获奖地点 (location)	地区
获奖时间 (time)	日期
奖金（prize amount）	金额

receive-Nobel-Prize是receive-prize的一个子类，其中的论元定义更为具体化。如：获奖内容即诺贝尔奖，获奖时间为年份，奖项分类为诺贝尔奖分设的所有六个奖项。

上述针对诺贝尔奖获奖关系的本体模型将被用于第3节的DARE实验。由于该领域模型

是基于通用本体构建的，其可扩展性因此得到了保障，并且可以和其它领域模型组合使用。基于SUMO和WorNet词义之间的对应关系，该领域模型可以直接访问WordNet查询词义。这对于使用词汇库进行推理，归纳并概括规则非常有利。

2.3 多元关系的定义

我们实验的应用领域涉及到多元的关系，包含两个或两个以上的论元。Krause等（2012）对该类多元关系给出了如下定义：

假设t 是命名实体的一种类别，则NE_t是t 类命名实体的集合。如果T是命名实体类别的一个多元组，$T = (t_1, t_2, \ldots, t_n)$，则在$T$上的一个$n$元关系$R$是一个满足如下条件的集合：

$$R \subseteq NE_{t_1} \times NE_{t_2} \times \cdots \times NE_{t_n}$$

任意一个属于这个集合的命名实体的多元组则被称为关系R的实例。

我们认为在同一个T上的所有关系属于同类关系。通常，关系的前k $(2 \leq k \leq n)$个论元是核心论元，关系是建立于这些论元实体之间的。因此，我们认为这些核心论元必须都出现在每个关系实例中。例如，在婚姻关系实例中我们要求结婚双方的人都必须被提及，而婚礼的时间和地点则属于可选论元。我们将所有的拥有同类核心论元的关系称为核心同类关系。 例如婚姻关系和父子关系就属于同类关系。

3 针对实用本体的高级信息抽取

通过关系抽取规则集或复杂关系抽取分类器，可以在语言模版和关系实例之间建立一种联系。这些关系由本体定义，其实例由命名实体的多元组构成，是一种简化的语义表达式。手工构建大型关系抽取规则的经验显示，这种方法费时耗力，并且覆盖率很低。因此，近年来关系抽取领域的研究愈加倾向于自动学习抽取规则或语言模板。关系抽取系统使用最多的是监督式学习方法，即从标注好的数据中学习。这些数据通常按照系统自动处理的期望结果和格式来进行标注（Muslea 1999）。这种方法获得了令人瞩目的成果。然而， 标注数据成本极其可观， 通常要求标注者具备一定的目标领域的专业知识。而且标注者之间的标注一致度也相对较低。另一方面，现有的非监督式学习对绝大多数的关系抽取任务来说并不合适。因为在没有知识背景的情况下，现有的数据并不能给出任何与目标关系相关的提示信息。与此相比，最小化的监督式迭代学习则更适合大部分关系抽取任务。这种方法把处理学习数据的人工成本降到了最低。无需任何语料标注，只通过若干样例，系统就可以学到全部关系抽取规则。最小化的监督式学习系统的初始化条件因系统而异，有些从抽取规则或者语言模板开始学习，有些则把语义样例或关系实例作为出发点。

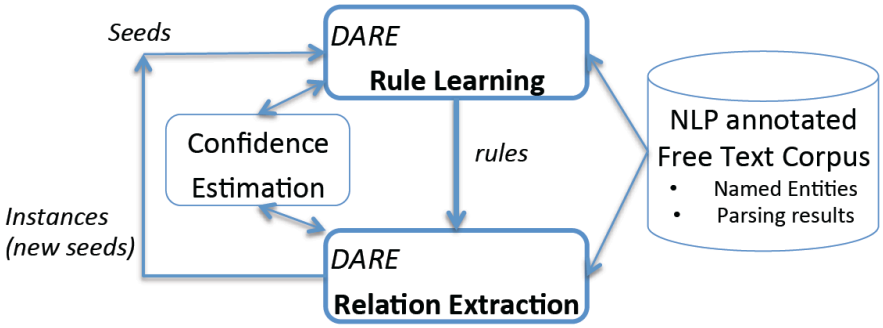
3.1 DARE 系统架构

DARE[4]是一个基于最小化的监督式机器学习系统，主要功能是从自然文本中学习语言模板和抽取关系实例。它能够使用其它语言处理系统，对语料进行预处理，即自动标注命名实体，分句和分析句法结构（如依存结构）。DARE 的关系处理核心主要由两

[4]英语：Domain Adaptive Relation Extraction，可适应领域关系抽取系。http://dare.dfki.de/

个部分组成：1）模板学习组件；2）关系抽取组件（Xu 2007；Xu，et al. 2007）。这两个组件相互依存构成了一个迭代（Bootstrapping）机器学习的框架。 机器学习从被称为“语义种子”的少数目标关系实例开始。首先， 学习组件从自动标注好的句子中，提取出和种子匹配的语言模板。关系抽取组件则进一步在更多的文本中应用这些模板来获取更多的关系实例。这些新的关系实例将被作为新的种子用于下一轮的学习。图 3.1 展示了 DARE 系统的总体架构。当不再发现新的学习模板或关系实例时，迭代循环将停止。由于学习仅仅依赖微量的领域知识——少数的几个语义种子，DARE 能够轻松地应对新领域中的其它关系类别。

图 3.1 DARE 的系统构架



针对多元关系，为了抽取其论元数目不同的实例，DARE 使用了一种递归表达方式定义多元的组合型模板，通过自下而上策略学习并组合这些模板。对于一个 n 元关系，一个 DARE 的学习模板可被分解成多条小型简单模板，每个都包含$k\ (1 \le k < n)$个论元。并且，DARE 的语言模板明确定义了那些语言学论元在目标关系中的语义角色。

我们将以获奖关系为目标关系， 来详细介绍 DARE 的语言模板学习机制。按照 Xu 等（2006）以及第 2 节的定义， 获奖关系描述了个人或组织在某一年获得某个类别的某种奖项的事件，即第 2 节简单本体中定义的 receive-prize 事件。我们的目标关系一共包含四个论元： 获奖者（recipient）、奖项（prize）、奖项分类（area） 和年份（year）。（马夫兹，诺贝尔奖，文学，1988）是一个以多元组的形式表达的获奖实例。这个实例可以用于描述如下例句中提及的获奖事件：

马夫兹曾于 1988 年荣获诺贝尔文学奖，成为第一位获此殊荣的阿拉伯作家。

图 3.2 显示了这个例句与获奖事件相关部分的依存结构。 从这个结构中，DARE 可以学习到图 3.3 和图 3.4 所显示的两个语言模板。图 3.3 中的一元模板可以从介词短语中提取语义论元年份(year)。图 3.4 显示的三元模板，可以从动词“荣获”支配的短语中提取出关系实例所包含的全部信息。“荣获”的主语是获奖者，宾语是一个复合名词短语形式的命名实体，其中包含了奖项和分类信息。同时这个三元的模板还调用了图 3.3 的二元模板，从补语中获取嵌套的时间论元。

图 3.2 例句与获奖事件相关的依存结构

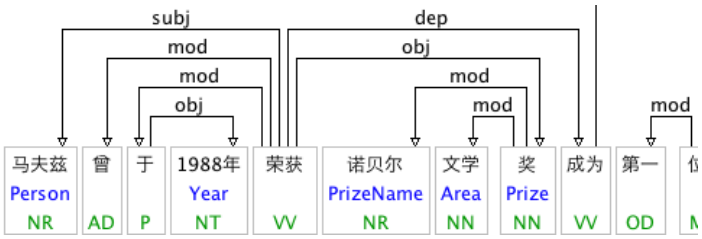


图 3.3 从图 3.2 种学习到的一元模板

模板名称：year1

模板主体：
$$\begin{bmatrix} \text{head} & \text{"于"，pos: P} \\ \text{obj} & [\text{head} \quad \text{Year}\boxed{1}] \end{bmatrix}$$

输出：⟨年份$\boxed{1}$⟩

图 3.4 从 3.2 中学习到的三元模板

模板名称：recipient_winner_prize_year

模板主体：
$$\begin{bmatrix} \text{head} & \text{"荣获"，pos: VV} \\ \textit{mod} & \textit{模板year}1\text{: ⟨年份}\boxed{1}\text{⟩} \\ \textit{subj} & [\text{head} \quad \text{Person}\boxed{2}] \\ \text{obj} & [\text{head} \quad \text{Prize⟨PrizeName}\boxed{3}\text{，Area}\boxed{4}\text{⟩}] \end{bmatrix}$$

输出：⟨年份$\boxed{1}$，获奖者$\boxed{2}$，奖项$\boxed{3}$，奖项分类$\boxed{4}$⟩

DARE 系统的核心组件通过非统计的机器学习方法，针对某个关系实例的多元组获取语言模版，然后重复使用这些模版获得新的多元组。对于一些关系类别和学习语料，使用这种方法已经能够很好的完成关系抽取任务。在这个基础上，我们还在现实应用中进一步扩展了 DARE 系统。如设置了模板和实例的置信度评估组件，来对获取的关系实例和抽取模版进行多方面的权衡。通过设置阈值，我们可以滤除低置信度的实例和模板，从而确保关系抽取的正确率。

3.2 英语新闻文本的关系抽取实验

我们以诺贝尔奖获奖和管理更替两个关系为目标来展示基于 DARE 系统的关系抽取。选取诺贝尔奖获奖关系作为目标关系的原因是该领域的事件有详尽的结构化数据记录，并且在互联网上可以获得大量报道这些得奖事件的自然语言文本。这些数据规模适度，相关的文本具有权威性并能被用于构造标准标注。这些都有利于不同种子的选取和测评。我们选用的文本来源于 New York Times，BBC 在线，CNN 新闻报道，及诺贝尔电子博物馆[5]的报道。抽取的目标关系是前文中提到的四元组：（获奖者，奖项，奖项分

[5] http://www.nobel.se/

类，年份）。管理更替是由 MUC-6（MUC-6 1995）定义的关系。我们使用 MUC-6 提供的测试文本，并进一步定义了相对简化的关系结构。这个目标关系同样包含四个论元（以便与诺奖关系的抽取结果进行比较）：personIn、personOut、position 和 organization。

DARE 机器学习开始之前，实验语料经过了其它自然语言处理系统的标注。我们首先利用 SProUT（Drozdzynski，et al. 2004）对语料进行命名实体标注。这一过程根据需要也可使用其它的命名实体识别系统。然后，我们使用 MINIPAR（Lin 2003）标注句子的依存结构。同样，这一标注也可换由其他句法分析器完成。

表 3.1 给出了针对两个关系的测试语料集的文章数量和大小。针对诺贝尔奖获奖关系，DARE 从一个种子开始学习语言模板和抽取关系实例，表 3.2 显现了抽取的结果。对于管理更替关系，DARE 基于 MUC-6 语料以不同数目种子分别进行学习，表 3.3 列出了这些实验的评测结果。

表 3.1 测试语料一览

数据名称	文章数量	数据量
诺贝尔奖获奖语料[6]	3300	20 MB
MUC-6 测试语料	199	1 MB

表 3.2 由一个种子开始的在诺贝尔奖语料上的 DARE 实验结果

目标关系	种子数量	正确率	召回率
诺贝尔奖获奖关系	1	80.59%	69%

表 3.3 由不同种子分别开始的在 MUC-6 语料上的 DARE 实验结果

目标关系	种子数量	正确率	召回率
管理更替关系	1	15.1%	21.8%
	20	48.4%	34.2%
	55	62.0%	48.0%

从表 3.1 可见，两个领域的语料之间存在很大差别，这就可以很好的解释实验结果的差距。获奖语料中语言模版和实例之间的关联度更接近于小世界网络（Small-world network）的特性（Amaral, et al. 2002）。因此，从单个实例出发的学习结果非常令人满意。而 MUC-6 语料具备的特性则不同，绝大多数情况下模版和实例之间不存在关联。因此，MUC-6 语料并不适合 DARE 的应用。Uszkoreit（2011）详细系统地描述了语料特性对于 DARE 实验的主要影响。

[6] 诺贝尔奖获奖关系的学习语料可以从 http://dare.dfki.de 下载，请联系 dare@dfki.de

4 中文语料的关系抽取实验

4.1 目标关系

我们调整了 DARE 系统，使其能够从中文的语料中抽取“诺贝尔奖获奖关系”（Xu，2007）的实例。按照 Xu 等（2006）以及第 2 节的定义，获奖关系是一种四元的目标关系，共包含四个论元：获奖者、奖项、奖项分类 和年份。我们把其中奖项的范围固定在诺贝尔奖，奖项分类则包括所有诺贝尔奖的领域：物理、化学、生理学或医学、文学、和平和经济。汉语中的诺贝尔奖的奖项分类和英文中的不同，它通常只会作为复合名词中的一部分出现，如：诺贝尔和平奖、诺贝尔物理奖等等。 而英文中除了 Nobel Chemistry Prize 则还会有 Nobel Prize in chemistry 这种表达方式，这时 chemistry 是一个单独的命名实体。 汉语中奖项分类的信息一般都被包含在奖的命名实体中，领域信息等极少单独作为命名实体出现。 所以和英文不同的是，DARE 从中文的语料中基本只会学习到二元或三元的语言模板，而不会学习到四元的模板。但是使用这些模板，仍然更够抽取到四元的关系实例。

4.2 实验语料和预处理

我们以中文 Gigaword 语料作为实验数据。从中文 Gigaword 语料的新华社文章中，我们抽取出了 2252 个包含了正则表达式“诺贝尔*奖”的句子，并利用这个表达式标注奖项和分类的信息。 这些句子被随机的分成两个部分： 1800 句作为 DARE 最小化监督式机器学习的语料，用来学习语言模板；其余的 452 句用来评测学习的结果。我们人工标注了所有句子中的所有人名和年份 （不包含人称代词）。从中我们选取了至少含一个人名的句子，并用 Stanford Parser 中支持分词的 xinhuaFactoredSegmenting.ser.gz 模型（Chang，et al. 2009 ）来分析这些句子的关系结构。该因子化（Factored）句法分析模型结合了短语结构和依存结构各有的优势，对候选句法结构进行统一评分。该句法分析模型同样是基于新闻类文本训练而得的，从而减少了可能因跨领域导致的分析错误。

我们在学习语料上用少量的初始知识进行 DARE 迭代自动学习的实验。 实验中学习到的语言抽取模板将应用到评测语料中来抽取关系实例。为了评测抽取的结果， 我们人工标注了所有评测语料中的诺贝尔奖获奖关系的实例。

4.3 DARE 迭代自动学习的实验

根据 Xu 等（2006）在英语语料上的 DARE 实验结果和 Uszkoreit（2011）的分析，初始种子的数量和质量对 DARE 迭代自动机器学习的效果有很大的影响。在针对中文学习语料的实验中， 我们也相应的选取了三组不同的种子来初始化 DARE 的机器学习。它们分别是：

1. 学习语料的句子中提到次数最少的一个人名，奖和年份的组合
   （艾伯特·卢图利，诺贝尔奖， 和平，1960）
2. 学习语料的句子中提到次数最多的一个人名，奖和年份的组合
   （佩雷斯， 诺贝尔奖， 和平，1994）
3. 学习语料的句子中提到次数最多的十个人名，奖和年份的组合
   （佩雷斯， 诺贝尔奖，和平，1994）

（拉宾，　诺贝尔奖，和平，1994）
（朱棣文，　诺贝尔奖，物理，1997）
（阿拉法特，诺贝尔奖，和平，1994）
（门楚，　诺贝尔奖，和平，1992）
（野依良治，诺贝尔奖，化学，2001）
（李政道，　诺贝尔奖，物理，1957）
（安南，　诺贝尔奖，和平，2001）
（杨振宁，　诺贝尔奖，物理，1997）
（白川英树，诺贝尔奖，化学，2000）

表 4.1 是分别以这三组种子开始的 DARE 规则迭代学习实验的详细数据结果：

表 4.1 不同种子开始的 DARE 在学习语料种运行的实验结果

实验	种子数量	DARE 运行中迭代的次数	学习到的语言模板数量		
			二元模板	三元模板	总量
1	1	1	1	1	2
2	1	6	533	89	622
3	10	5	533	131	664

从表 4.1 的结果可看出，同是由一个种子开始的实验，实验 1 中用一个种子只能学习到很少的语言模板，并且利用这些模板并不能得到新的关系实例从而继续学习模板； 而实验 2 则能真正开始迭代地学习模板。这说明一个“好的”种子对 DARE 的实验是非常重要的。同时，实验 3 表明，如果能够以更多的种子开始，则能够学习到更多的模板。 实验 3 中增加的三元模板说明，更多的初始信息，则能够提高学习到的模板的质量。

5 评测结果与分析

5.1 评测结果

我们把从第四节三个实验中学习到的语言模板应用到评测语料中，通过评测抽取出的关系实例来评测这些模板的质量。对于从句子中抽取到的每一个二元或多元的关系实例，我们都会对比人工标注来评测实例中的命名实体是否正确。表5.1是相应的不同实验的抽取结果的评测数据。

表5.1 使用从不同实验中学习到的模板从评测语料中抽取出的关系实例的评测

模板来源	抽取的关系实例中的命名实体总量	正确率	召回率	F1
实验1	0	-	0	-
实验2	316	91.46%	35.68%	51.33%
实验3	321	90.97%	35.83%	51.41%

实验1中学习到的少量模板，并不能应用到评测语料中。相比之下，实验2和实验3中学习到的模板，质量很高，并能够从评测语料种抽取到一定量的关系实例。为了能更好的了解DARE的实验结果，我们把所有在同一句子种出现的人名，奖和年份的多元组合

作为默认的关系实例进行评测。其召回率是100%，准确率是71.24%。 这说明，利用DARE，我们能更精确的抽取到关系实例。

表5.2详细介绍了在实验2和实验3中学习到的模板在评测语料中的应用细节。

表5.2 实验2和实验3中学习到的模板在评测语料种的应用细节

规则来源	抽取中使用到的模板			好的模板			危险的模板			坏的模板
	二元	三元	总量	二元	三元	总量	二元	三元	总量	
实验2	46	8	54	34	6	40	12	2	14	0
实验3	45	9	54	33	7	40	12	2	14	0

根据Xu等（2010）的定义，所谓好的模板就是指使用后只抽取到正确的关系实例的模版；而与之相对， 坏的模板只能抽取到错误的关系实例。如果一个模版既能得到正确的关系实例，也能得到错误的，这个模板就是危险的。表5.2表明，虽然实验2和实验3中都有54条规则能够运用在评测语料中，但是实验3的模板则包含了更多的信息，从而能得到更高的召回率。 根据抽取到的结果的正确性，我们计算了不同元模板的抽取正确率（见表5.3）。和在英语语料中的实验相同，三元的模板的质量，要略高于二元模板。

表5.3 实验2和实验3中学习到的不同元的模板的质量分析

	实验2的模板		实验3的模板	
	二元模板	三元模板	二元模板	三元模板
抽取结果的正确率	90.65%	92.59%	90%	93%

5.2 评测结果分析

我们将以实验3的评测结果为例，分析DARE的中文关系抽取实验。表5.1可见，中文实验中抽取出部分错误的关系实列，并且抽取的召回率与英语实验相比偏低。这两个问题出现的原因主要有两点：1）学习语料过少；2）句子分析错误。我们将针对这两点原因在下两小节分别进行说明。

5.2.1 学习语料对召回率的影响

与英语实验相比，中文实验虽有较高的准确率，但是召回率偏低。召回率缺失的主要原因之一就是学习语料量过小。和英语学习语料相比，中文的学习语料包含的相关句子数量比较少。一些相关的获奖表达方式在学习语料种并没有出现，如

* *今年诺贝尔奖两得主之一的美国科学家阿格雷 ……*
* *与阿格雷分享诺贝尔化学奖的是另一名美国科学家麦金农 ……*
* *印度大诗人泰戈尔 1913 年为亚洲赢得的第一枚诺贝尔奖章日前……失窃*

解决这个问题的方法是扩大学习语料，例如以诺贝尔奖获奖名单作为关键字使用搜索引擎从网络收集语料 (Krause 2012)。 不过这一方法的弊端是收集的语料包含过多的不相关信息， 会使学习到的规则质量下降。

5.2.2 中文语言现象造成的错误分析

除了实验语料库规模有限导致的抽取实例数量不高外，中文特有的一些语言现象给 DARE 实验造成了一些困难，同时影响了正确率和召回率。这些困难并非是单独针对关系抽取任务而存在的。但是对于一个原本处理英文的系统而言，这些困难需要特殊的处理，有待将来进一步的研究。

* 中文分词的问题

  由于缺少明确的词边界，在中文分析中不可或缺的一个步骤是分词。虽然现有的分词技术在普通文本上的总体正确率已经很高，对于未知词的切分效果仍然不尽人意。而在诺奖领域，充斥着大量的外来语和专有名词，如外国人名、地名、组织名、及其各种缩略简称。这些未知词所对应的实体恰恰又是目标关系的核心论元。在错误分析中，我们发现许多错误都是由不正确的分词导致的。例如，以下句子中的多个日本人名被错误的切分，以致依存树混乱错误。

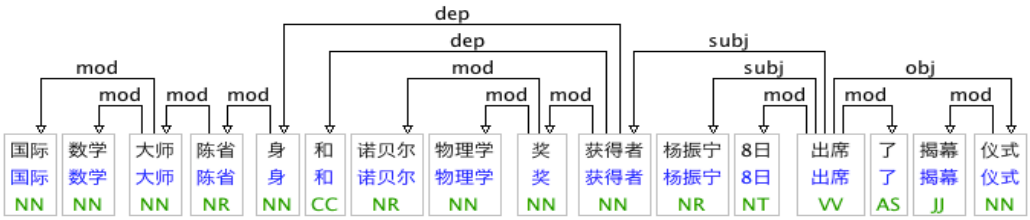
  野依良治是/继福井谦一和白川英树/之后，第三/位/获得/诺贝尔/化学/奖/的/日本/科学家/。

  这种情况会使 DARE 无法学习到正确有用的语言模板，亦或学到的好的模板并不能应用到测试语料中从而抽取出关系实例，进而导致了召回率的缺失。

* 复杂联合结构和同位结构给句法分析带来的困难

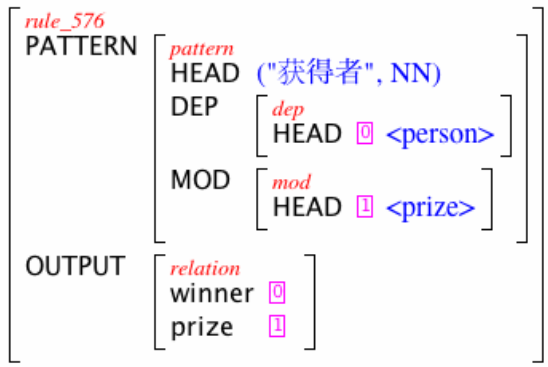
  我们发现在新闻报道的句子中，存在着大量的复杂联合结构和同位结构。从句法角度，不存在能够严格区分两者的严格约束。所以在句法分析的过程中，难免会发生无法正确消歧的情况。例如图 5.1 所示，"诺贝尔物理学奖获得者"与"杨振宁"之间的同位语关系没有被正确识别，导致相应的获奖关系无法通过已有的规则抽取。

图 5.1 错误的同位语分析实例



而错误的“获奖者”和“陈省身”之间的同位语关系，导致 DARE 使用图 5.2 所示的学习模板抽取出错误的关系实例，使实验的正确率的下降。

图 5.2 DARE 学习到的同位语结构的抽取模板



6 总结

本文介绍了本体和信息抽取技术之间的关系以及如何使用简单本体定义实用的信息抽取目标。以关系抽取系统 DARE 为例，本文详细描述了一个最小化监督式机器学习系统的工作流程和系统特性，以及抽取结果对初始化信息、学习语料等的依赖程度。DARE 在不同的英文和中文语料上的运行和评测结果表明，通过少量的语义种子，最小化监督式学习系统即可从相应的语料中学习到高质量的语言模板。这些模板可以应用到其它的语料中来抽取目标关系实例。这些实验同时也表明，预处理和句法分析的质量对 DARE 的学习和抽取有决定性的影响。DARE 的运行质量非常依赖于句法分析的结果是否准确或一致。为了提高中文抽取的召回率，在今后的实验中尝试使用其它的中文分词器或句法分析工具是非常必要的。Xu 等（2011）通过评测关系抽取系统的学习模板和实例来对英文句法分析的结果重新排序，这个方法同样也可以尝试使用到中文的实验中。提高中文实验的召回率的另外一个主要的手段是尽可能的扩大学习语料。今后，我们计划使用其它的大型文本作为学习语料进行实验，如 Wikipedia 等。另外，我们也将尝试定义一些其它的抽取目标关系， 并将中文和英文的抽取结果进行比较分析。

参考文献

Agichtein, E. and L. Gravano. 2000. *Snowball: extracting relations from large plain-text collections.* Proceedings of the fifth ACM conference on Digital libraries. Pp. 85 – 94. New York, NY, USA. ACM.

Amaral, L. A. N.,A. Scala, M. Barthélémy and H. E. Stanley. 2005. *Classes of small-world networks.* Proceedings of the National Academy of Sciences, 102(30). Pp. 10421 – 10426.

Appelt, D. 2003. *Semantics and information extraction.* Center for Language and Speech Processing.

Appelt, D. and D. Israel. 1999. *Introduction to information extraction technology.*

Berners-Lee, T. 1999. *Weaving the Web: The original design and ultimate destiny of the World Wide Web by its inventor.* New York: Harper Collins Publishers.

Chang, Pi-Chuan, Huihsin Tseng, Dan Jurafsky and Christopher D. Manning. 2009. *Discriminative Reordering with Chinese Grammatical Relations Features.* Third Workshop on Syntax and Structure in Statistical Translation.

Drozdzynski, W., H.-U. Krieger, J. Piskorski, U. Schäfer and F. Xu. 2004. *Shallow processing with unification and typed feature structures — foundations and applications.* Künstliche Intelligenz, 1:17 – 23.

Frank, A., H. Krieger, F. Xu, H. Uszkoreit, B. Crysmann, B. Jörg and U. Schäfer. 2006. *Question answering from structured knowledge sources.* Journal of Applied Logic, Special Issue on Questions and Answers: Theoretical and Applied Perspectives 1, 29.

Grishman, R. and B. Sundheim. 1996. *Message understanding conference - 6: A brief history.* Proceedings of COLING 1996, 466-471.

Hearst, M. 1992. *Automatic Acquisition of Hyponyms on Large Text Corpora.* Proceedings of the Fourteenth International Conference on Computational Linguistics.

Krause, S., H. Li, H. Uszkorei and F. Xu. 2012. *Learning Large-Scale Relation Extraction Rules with Distant Supervision from Web.* The 11th International Semantic Web Conference (ISWC 2012), Boston, USA, November 2012.

Lin, D. 2003. *Dependency-based evaluation of MINIPAR.* In Anne Abeill´e, editor, *Treebanks - Building and Using Parsed Corpora.* Kluwer Academic Publishers.

Miller, G. A., R. Beckwith, C. Fellbaum, D. Gross and K. Miller. 1993. *Five papers on WordNet.* Technical report, Cognitive Science Laboratory, Princeton.

MUC-6. 1995. *Proceedings of the 6th conference on message understanding.*

Muslea, I. 1999. *Extraction patterns for information extraction tasks: A survey.* The AAAI Workshop on Machine Learning for Information Extraction, Orlando, Florida.

Niles, I. and A. Pease. 2003. *Linking lexicons and ontologies: Mapping WordNet to the suggested upper merged ontology.* Proceedings of the 2003 International Conference on Information and Knowledge Engineering.

Niles, I. and A. Terry. 2004. *The MILO: A general-purpose, mid-level ontology.* The 2004 International Conference on Information and Knowledge Engineering, Las Vegas, NV.

Pease, A., I. Niles and J. Li. 2002. *The Suggested Upper Merged Ontology: A Large Ontology for the Semantic Web and its Applications.* Working Notes of the AAAI-2002 Workshop on Ontologies and the Semantic Web 28.

Ravichandran, D. and E. Hovy. 2002. *Learning surface text patterns for a question answering system.* Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, July 2002. Pp. 41 – 47.

Stevenson, M. and M. Greenwood. 2005. *A semantic approach to IE pattern induction.* Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL' 05), Michigan, June 2005. Pp. 379 – 386.

Uszkoreit, H. 2011. *Learning relation extraction grammars with minimal human intervention: Strategy, results, insights and plans.* In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing.* Proceedings of the 12th International Conference on Intelligent Text Processing and Computational Linguistics volume 6609. Lecture Notes in Computer Science, Springer Berlin / Heidelberg. Pp. 106 – 126.

Yangarber, R., R. Grishman and P. Tapanainen. 2000. *Automatic acquisition of domain knowledge for information extraction.* Proceedings of the 18th International Conference on Computational Linguistics, pages 940 – 946.

Xu, F., H. Uszkoreit and H. Li. 2006. *Automatic event and relation detection with seeds of varying complexity.* Proceedings of AAAI 2006 Workshop Event Extraction and Synthesis, Boston.

Xu, F. 2007. *Bootstrapping Relation Extraction from Semantic Seeds.* Ph.D. diss., Saarland University.

Xu, F., H. Uszkoreit and H. Li. 2007. *A seed driven bottom-up machine learning framework for extracting relations of various complexity.* Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL 2007), Prague, Czech Republic, June.

Xu, F., H. Uszkoreit, S. Krause and H. Li. 2010. *Boosting relation extraction with limited closed-world knowledge.* The 23rd International Conference on Computational Linguistics (COLING 2010), Posters, Beijing, China, August. Pp. 1354 – 1362.

Xu,F., Hong Li, Yi Zhang, Hans Uszkoreit and Sebastian Krause. 2011. *Minimally Supervised Domain-Adaptive Parse Reranking for Relation Extraction.* Proceedings of International Conference on Parsing Technologies (IWPT 2011), Dublin, Ireland, 2011.