

硕士学位论文

开放式中文实体关系抽取研究

RESEARCH ON CHINESE OPEN ENTITY RELATION EXTRACTION

刘安安

哈尔滨工业大学

2013 年 6 月

国内图书分类号：TP391.2

学校代码：10213

国际图书分类号：681.37

密级：公开

工程硕士学位论文

开放式中文实体关系抽取研究

硕 士 研 究 生：刘安安

导 师：秦兵教授

申 请 学 位：工程硕士

学 科：计算机技术

所 在 单 位：计算机科学与技术学院

答 辩 日 期：2013 年 6 月

授予学位单位：哈尔滨工业大学

Classified Index: TP391.2

U.D.C: 681.37

Dissertation for the Master Degree in Engineering

RESEARCH ON CHINESE OPEN ENTITY RELATION EXTRACTION

Candidate:	Liu Anan
Supervisor:	Prof.Qin Bing
Academic Degree Applied for:	Master of Engineering
Speciality:	Computer Technology
Affiliation:	School of Computer Science and Technology
Date of Defence:	June, 2013
Degree-Conferring-Institution:	Harbin Institute of Technology

摘要

实体关系是描述实体之间语义关系的重要途径。实体关系抽取是信息抽取任务中的重要环节，也有着广泛的应用前景。随着 Web2.0 的迅猛发展，人们对实体关系抽取提出了新的要求，以适应从快速增长的海量互联网文本中迅速准确地获取对用户有价值的信息。

传统的实体关系抽取需要预先定义关系类型体系，然而定义一个全面的实体关系类型体系是很困难的。开放式实体关系抽取技术通过使用关系指示词描述关系的方法解决了预先定义关系类型体系的问题，但是在中文上的研究还比较少。因此，针对不同的应用场景，本文提出了两种不同的开放式实体关系抽取方法，并且探索自动构建关系类型体系的相关方法。

针对句子的开放式实体关系抽取问题，本文提出基于有指导的开放式实体关系抽取方法。首先，制定开放式实体关系抽取语料标注规范，并且构建开放式实体关系抽取语料库；然后，通过分析语料中的语言现象，制定了先识别实体对和先识别关系指示词两套方法，并且设计了泛化能力强的特征抽取方案。在开放式实体关系语料上测试的 F 值达到 61.41%。

针对互联网的开放式实体关系抽取问题，本文提出面向大规模网络文本的无指导开放式中文实体关系抽取（UnCORE: Unsupervised Chinese Open Entity Relation Extraction for the Web）方法，首先使用实体之间的距离限制和关系指示词的位置限制获取候选关系三元组，然后采用全局排序和类型排序的方法来挖掘关系指示词，最后使用关系指示词和句式规则对候选关系三元组进行过滤得到最终的关系三元组。在获取大量关系三元组的同时，还保证了 80% 以上的微观平均准确率，满足实用要求。

本文使用基于关系指示词聚类的方法自动构建关系类型体系。基于 RNN-LM 的关系指示词相似度计算方法和基于 HowNet 的关系指示词相似度计算方法，尝试使用近邻传播聚类算法和层次聚类算法对关系指示词进行聚类。在 PER-PER 实体对类型的关系指示词集合上实验，平均 F 值最高达到 64.25%。

最后，为了把本文的相关研究成果展示给研究人员，搭建了两个演示系统：面向句子的开放式实体关系抽取系统和开放式实体关系三元组查询系统。面向句子的开放式实体关系抽取系统从用户输入的自然语言文本中抽取关系三元组，并且把抽取结果以网页的形式展现。开放式实体关系三元组查询系统对从互联网文本中挖掘的大量关系三元组构建索引，用户输入查询的实体，系统返回该实体相关

的关系三元组，并且以清晰直观的方式展示关系三元组。

关键词 开放式；实体关系抽取；关系三元组；关系指示词；关系类型体系

Abstract

Entity relationship is an important way to describe the semantic relationships between entities. As one of the most important subtask of information extraction, entity relation extraction has wide application prospects. With the rapid development of the Web2.0, people put forward new requirements on the entity relation extraction to accommodate quickly and accurately obtaining valuable information on the rapid growth of massive web text for user.

Traditionally, Entity Relation Extraction (RE) methods required a pre-defined set of relation types. But it's difficult to build a well-defined architecture of the relation types. Open Entity Relation Extraction (ORE) is the task of extracting relation triples from natural language text without pre-defined relation types. We propose two ORE methods to solve relation extraction on different application scenarios, and explore solutions to automatically build relation types.

This paper presents a supervised method to solve sentence-level ORE problem. The detailed criterion of annotation is established and a corpus which contains 1000 documents is annotated. By analyzing the linguistic phenomenon of the corpus, we design a domain-independent program to extract features. The average F-measure achieves 61.64% on the corpus.

This paper presents UnCORE (Unsupervised Chinese Open Entity Relation Extraction for the Web), an unsupervised ORE method which is to discover relation triples from large-scale web text. UnCORE exploits word distance and entity distance constraints to generate candidate relation triples, and then adopts global ranking and domain ranking methods to discover relation words from the relation triple candidate. Finally UnCORE filters them by using the extracted relation words and some sentence rules. Results show that UnCORE extracts large scale relation triples at precision higher than 80%.

This paper proposes the relation-words-clustering-based method to build the relation types. First, we calculate the similarity between relation words based on RNN-LM or HowNet, and then cluster the relation words by AP or HAC. Finally, we build a well-defined relation types.

At last, we design and implement a demonstration platform for users to extract relation triples from sentence and to search relation triple.

Keywords Entity Relation Extraction, Relation Triple; Relation Word, Relation Types

目 录

摘 要	I
Abstract	III
第 1 章 绪 论	1
1.1 课题来源	1
1.2 课题背景	1
1.3 研究目的和意义	2
1.4 关系抽取的研究现状	2
1.4.1 预测实体对之间的关系	3
1.4.2 挖掘特定关系的实体对	5
1.4.3 开放式实体关系抽取	7
1.5 本文的主要研究内容	12
第 2 章 面向句子的开放式中文实体关系抽取	14
2.1 引言	14
2.2 语料构建	14
2.3 有指导的开放式实体关系抽取	18
2.3.1 语料分析	18
2.3.2 先识别实体对的方法	20
2.3.3 先识别关系指示词的方法	21
2.4 实验结果及其分析	22
2.4.1 先识别实体对的方法	22
2.4.2 先识别关系指示词的方法	24
2.4.3 实验结果分析	24
2.5 本章小结	25
第 3 章 面向互联网的开放式中文实体关系抽取	26
3.1 引言	26
3.2 无指导的开放式实体关系抽取	26
3.2.1 预处理	27
3.2.2 生成候选三元组	28
3.2.3 生成关系指示词词表	30
3.2.4 后处理	31
3.3 实验结果及其分析	33

3.3.1 数据及评价方法	33
3.3.2 结果及分析	34
3.3.3 错误分析	39
3.4 本章小结	40
第 4 章 开放式中文实体关系类型体系自动构建	42
4.1 引言	42
4.2 基于聚类的开放式实体关系类型体系自动构建	42
4.2.1 相似度计算	43
4.2.2 聚类	43
4.3 实验结果及其分析	45
4.3.1 数据与评价标准	45
4.3.2 结果与分析	47
4.4 本章小结	49
第 5 章 开放式中文实体关系抽取平台设计与实现	51
5.1 引言	51
5.2 句子级开放式实体关系抽取系统	51
5.3 开放式实体关系三元组查询系统	53
5.4 本章小结	54
结 论	55
参考文献	57
攻读硕士学位期间发表的论文及其它成果	61
哈尔滨工业大学学位论文原创性声明和使用权限	62
致 谢	63

Contents

Abstract (In Chinese)	I
Abstract (In English)	III
Chapter 1 Introduction	1
1.1 Source of subject	1
1.2 Background of the subject	1
1.3 Objective and significance of the subject	2
1.4 Research status of Relation Extraction	2
1.4.1 Predict the relationship between entities	3
1.4.2 Mining instance with specific relationship	5
1.4.3 Open Entity Relation Extraction	7
1.5 Main research contents of this subject	12
Chapter 2 Open Entity Relation Extraction for sentence	14
2.1 Introduction	14
2.2 Open Entity Relation Extraction Corpus building	14
2.3 Supervised Method for Open Entity Relation Extraction	18
2.3.1 Corpus analysis	18
2.3.2 Identify the entity pair first	20
2.3.3 Identify the relation words first	21
2.4 Experiment and analysis	22
2.4.1 Identify the entity pair first	22
2.4.2 Identify the relation words first	24
2.4.3 Result analysis	24
2.5 Summary	25
Chapter 3 Open Entity Relation Extraction for web	26
3.1 Introduction	26
3.2 Unsupervised Method for Open Entity Relation Extraction	26
3.2.1 Preprocessing	27
3.2.2 Generate candidate triples	28
3.2.3 Generate relation words	30
3.2.4 Post-processing	31
3.3 Experiment and analysis	33

3.3.1 Evaluation Method	33
3.3.2 Results and analysis	34
3.3.3 Error analysis	39
3.4 Summary	40
Chapter 4 Building set of open entity relation types	42
4.1 Introduction	42
4.2 Building set of open entity relation types based on clustering	42
4.2.1 Calculating similarity	43
4.2.2 Clustering	43
4.3 Experiment and analysis	45
4.3.1 Evaluation Method	45
4.3.2 Results and analysis	47
4.4 Summary	49
Chapter 5 Open entity relation extraction platform design and implementation .	51
5.1 Introduction	51
5.2 Open Entity Relation Extraction for sentence	51
5.3 Open Entity Relation Extraction for web	53
5.4 Summary	54
Conclusion	55
References	57
Papers published in the period of master education	61
Statement of copyright and Letter of authorization	62
Acknowledgements	63

第1章 绪 论

1.1 课题来源

本课题的主要研究内容来自于国家重点自然科学基金项目《中文篇章级语义分析理论与方法》。

1.2 课题背景

随着 Web2.0 的兴起，互联网的普通用户可以参与到互联网的建设中来，向互联网贡献着大量的文本数据，例如社交网站、微博、博客、贴吧、论坛、百科知识等等。互联网的快速增长为人们提供了一个取之不尽用之不竭的信息源，怎么使用一种自动的方法对这些文本进行处理，快速准确地从海量文本中抽取对用户有用的信息成为人们关注的焦点。

搜索引擎（Search Engine）在一定程度上解决了用户的问题。搜索引擎（如谷歌、百度）对网页结构进行连接分析计算网页的重要性返回和查询相关的网页，这种方法能获取用户感兴趣的网页，但是无法对文本进行深层次的理解，所以还需用户浏览大量网页获取有用的知识。对文本进行深层分析可以为用户提供更加精准的服务。

信息抽取（Information Extraction, IE）技术正是在这种背景下产生。信息抽取的主要目的是从自然语言文本中抽取指定的实体（Entity）、关系（Relation）、事件（Event）等事实信息。信息抽取技术可以把文本中蕴含的无结构化信息转化成结构化的信息，存储在数据库中，方便用户检索，快速获取感兴趣的信息。

实体关系抽取（Entity Relation Extraction）是信息抽取的子任务，其主要目的是识别实体之间的语义关系。在传统的关系抽取任务中，需要预先定义好关系类型体系。例如雇佣关系（employee-of）、整体部分关系（part-whole）、位置关系（location）等等。文本“…百度董事长兼首席执行官李彦宏…”中的“百度”（机构）和李彦宏（人物）两个实体之间构成雇佣关系（employee-of），即“李彦宏”受雇于“百度”。通过以上介绍可以发现，如果说信息抽取的主要功能是自动将非结构化的自然语言文本表述为结构化的表格数据，实体识别确定了表格中各个元素的话，那么实体关系抽取则是确定这些元素在表格中的相对位置^[1]。由此可见，实体关系抽取的主要目的是在实体识别的基础上，把无结构的自然语言文本中所

蕴含的实体之间的语义关系抽取出来，整理成结构化的三元组（关系，实体 1，实体 2）存储在数据库中，供用户查询或者进一步分析利用，是信息抽取中非常重要的一个任务。

1.3 研究目的和意义

传统的实体关系抽取方法需要预先定义实体关系类型体系，针对预先定义好的每一类实体关系人工标注训练语料，然后利用机器学习的方法训练分类器进行新的关系实例识别和关系元组抽取。然而，预先定义一个全面实体关系类型体系是很困难的，并且人工构建大规模的语料库是及其耗时耗力的。所以，急需一种无指导的方法，可以自动的完成关系类型发现和关系抽取任务，避免预先定义关系类型体系和人工构建语料库。

开放式实体关系抽取技术^[8]应运而生，开放式实体关系抽取技术使用实体对上下文的一些词语来描述实体之间的语义关系，从而避免了构建关系类型体系。开放式实体关系抽取任务是在文本中抽取关系三元组（entity1，relationWords，entity2），其中（entity1，entity2）是存在关系的实体对，relationWords（关系指示词）是上下文中描述实体对的语义关系的词或词序列。在文本“早在去年，腾讯首席执行官马化腾就多次全面阐述了腾讯的发展战略。”中可以抽取关系三元组（腾讯，首席执行官，马化腾）。目前中文上的开放式实体关系研究还比较少，也没有公共的评价体系。本文在提出中文上的开放式实体关系抽取方法的同时，还将构建一个标准的开放式中文实体关系抽取评价语料。

将本文的方法应用到大规模语料上，可以抽取大量的关系三元组，这些关系三元组可以应用到下列任务中：

- 1) 构建知识图谱。通过知识图谱，搜索引擎可以给用户呈现出更加精准的信息，其搜索结果是知识而不是普通文本；
- 2) 问答系统。为问答系统提供大规模结构化信息，当用户提问“美国总统有哪些人？”，将从关系三元组中找到结果。

1.4 关系抽取的研究现状

实体关系抽取技术已经被广泛应用到自然语言文本中，包括新闻^[2]、科学出版物^[3]、博客、电子邮件^[4]、维基百科^{[5][6]}和普通的网络文本^{[7][8]}。MUC（Message Understanding Conference）^{[9][10]}和 ACE（Automatic Content Extraction）^[2]评测会议促进了关系抽取研究的蓬勃发展。

MUC 会议由美国国防高级研究计划委员会 (Defense Advanced Research Projects Agency, DARPA) 资助, 在 1987 年到 1998 年共召开了七届。关系抽取任务于 1998 年在 MUC-7^[11]正式提出, 其任务是确定实体之间的语义关系^[12]。

在 MUC-7 之后, MUC 被美国国家标注技术研究院 (National Institute of Standards and Technology, NIST) 引导的 ACE 评测会议所取代。ACE 评测会议从 1999 年到 2008 年至今共举办过八届, 每次评测会议的有所不同。最近一届 ACE 评测会议是 ACE08, 于 2008 年 5 月举办。ACE08 的关系抽取任务共定义了 7 个大类、18 个子类^[13]。

现阶段的关系抽取研究可以分为三个方向:

- 1) 预测实体对之间的关系
- 2) 挖掘特定关系的实体对
- 3) 开放式实体关系抽取

1.4.1 预测实体对之间的关系

这个任务的主要目的是, 给定一个关系类型体系 R , 根据两个实体的上下文, 预测这两个实体的语义关系。这两个实体往往出现在同一个句子当中, 所以这个任务可以这么描述: 给定一个句子 s 以及 s 中的两个实体 E_1 和 E_2 , 预测 E_1 和 E_2 在句子 s 中的关系类型 rel , rel 的候选集合是 R 。例如 (美国, 奥巴马) 在句子“奥巴马当选美国总统”中是“雇佣”关系, 而在句子“奥巴马出生于美国”中是“籍贯-出生地”关系。目前有主要有三类方法: 基于规则 (Rule-based Methods) 的方法、基于特征抽取的方法 (Feature-based Methods) 和基于核函数的方法 (Kernel-based Method)。

1) 基于规则的方法

基于规则的方法需要书写描述两个实体所在结构的规则,^{[14][15][4][16]}描述了一系列的基于规则的实体关系抽取系统。这种方法要求规则构建者都领域的特点有深入的了解, 投入成本大, 移植性差, 所以逐渐被其他方法所取代。

2) 基于特征的方法

从实体的上下文、词性、句法等信息中抽取特征训练一个分类器 (决策树、最大熵、支撑向量机等), 从而完成关系抽取任务。Jiang 等人^[17]提出了一套系统的方法从各种信息中抽取特征。下面将具体介绍抽取句子“哈尔滨工业大学校长王树国”的特征。

➤ 浅层特征: 分词和命名实体识别后的结果为“国务院\Organization 总理\n 温家

宝\Person”，“国务院”和“温家宝”是句子中的两个实体。每个词 x_i 有若干属性，比如词本身、实体类别、词性等。对于词“国务院”的一元特征有：

(文本=国务院)
(词性=“n”)
(实体类型=“Organization”)

二元特征有：

(文本=“国务院_总理”)
(词性=“n_n”)
(实体类型=“Organization_null”)

使用这种方法可以抽取大量的特征。

- 深层特征：上述句子的依存句法分析结构如图 1-1 所示。在这类方法中，一般先找出两个实体的最短路径，然后以依存弧为节点抽取特征。在图 1-1 中，最短依存路径是“国务院←总理←温家宝”，对第一条依存弧抽取一元特征：

(依存弧=“ATT”，文本=“国务院_总理”，词性=“n_n”)

同样地，可以抽取二元特征：

(依存弧=“ATT_ATT”，文本=“国务院_总理_温家宝”)

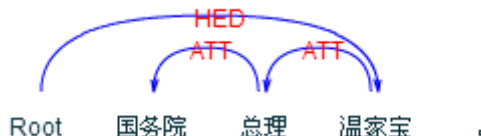


图 1-1 “国务院总理温家宝”的依存句法分析结果

Fig. 1-1 Dependency path for sentence “国务院总理温家宝”

在训练语料中可以统计出大量的特征，从而导致维数灾难问题。董静等人^[18]对语料库的特点进行分析，提出将实体关系划分为包含实体关系与非包含实体关系两类，同一种句法特征在这两类问题中的作用有明显的差异，从而选取不同句法特征集合，同时提出适合各自特点的新特征。在 CRF 模型下，以 ACE2007 语料作为实验数据，结果表明划分方法和新特征有效提高了汉语实体关系抽取任务的性能。Sun 等^[19]提出特诊稀疏现象对实体关系抽取任务的性能有很大的影响。

3) 基于核函数的方法

基于特征的方法需要考虑如何将非线性结构（句法）转换成线性结构，而基于核函数的方法不用这样做。基于核函数的方法可以利用核函数直接计算两个非线性结构的相似度，不需要抽取特征，从而也没有维数灾难问题。基于核函数的方法最重要的一步是设计一个计算两个实例(X, X')相似度的核函数(Kernel function) $K(X, X')$ 。基于核函数的方法最早在 SVM 模型中引入。

设训练实例 (x^i, E_1^i, E_2^i, r^i) ，其中 $i = 1 \dots N$ ， N 是语料库中训练实例的数目。我们使用 X_i 来代表 (x^i, E_1^i, E_2^i) ，使用公式 (1-1) 预测 $X(x, E_1, E_2)$ 的关系类别 \hat{r} 。

$$\hat{r} = \arg \max_r \sum_{i=1}^N \alpha_{ir} K(X_i, X) \quad (1-1)$$

α_{ir} 是第 i 个训练实例在类别 r 上的权重，需要在模型训练的时候估计其取值。

Tsochantaridis 等人^[20]对训练方法有详细的研究。

Bunescu 等人^{[21][22][23][24][25][26]}对短语句法和依存句法上的核函数有深入的研究，在依存句法上的核函数研究更加广泛。[21]提出了计最短依存路基相似度的核函数，设 T 是关系实例 $X(x, E_1, E_2)$ 两个实体的最短依存路径， $\{P\}$ 是依存路径上的节点集合， $\{p\}$ 是各个节点的属性集合，核函数 $K(X, X')$ 的计算公式如 (1-2) 所示。

$$K(X, X') = \begin{cases} \lambda \prod_{k=1}^{|\{P\}|} \text{CommonProperties}(P_k, P'_k) & \text{if } |\{P\}| = |\{P'\}| \\ 0 & \text{otherwise} \end{cases} \quad (1-2)$$

$\text{CommonProperties}(P_k, P'_k)$ 是节点 P_k 和节点 P'_k 中相同属性值的个数。从公式 (1-2) 中不难发现，当 T 和 T' 的长度相同并且路径上节点的属性相似的时候，核函数会得到一个比较大的数值。公式 (1-2) 有一个很大的缺点是，一旦依存路径的长度不相等的时候，核函数的值为 0。使用卷积核函数计算字符串、短语结构句法和依存句法的相似度，可以避免公式 (1-2) 的缺点。卷积核函数最初使用于字符串的相似度中，但是被扩展应用到树状结构。Zhang 等人^[25]的研究表明，在关系抽取任务中使用卷积核函数可以得到更好的性能。

1.4.2 挖掘特定关系的实体对

不同于前一任务，挖掘满足特定关系的实体对的主要目的是在大规模语料中抽取出满某一种特定关系的实体对^{[27][7][8][28][29][30][31]}。由于要处理大规模语料库，所以对系统的速度要求很高。近期的工作主要集中在挖掘大规模网络文本中的关系实例，使用半指导的方法，不需要标注语料，只需人工给定少量的关系种子。该方法的输入有如下几类：

- 构成关系的实体类型。如雇佣关系是由“机构”和“人物”两类实体构成的，

即 employee-of (Organization, Person);

- 关系种子集合 S 。关系种子是指能描述关系，例如 employee-of (哈尔滨工业大学, 王树国)。同时需要不能描述该关系的种子，用来生成反例;
- 人工书写的模板，这类输入是可选的。例如雇佣关系的一个模板 ([Organization] 校长 [Person])，又如籍贯关系的一个模板 ([Person] 出生于 [Location]);

给定输入后，有三个主要的步骤来解决问题。给定语料库 D ，关系类型 r ，构成关系 r 的实体类型对类型 (E_1, E_2) ，关系种子集合 S (包含正例 P^+ 和反例 P^-)。雇佣关系的关系种子集合如表 1-1 所示

表 1-1 雇佣关系的种子

Table 1-1 Relation seeds of employee-of

实体 1	实体 2	极性
哈尔滨工业大学	王树国	正例
国务院	温家宝	正例
百度	比尔盖茨	反例

第一步，通过种子集合 S 学习出模板集合 M 。第二步，使用模板集合在生语料库 D 中挖掘候选的关系实例 (r, E_1, E_2) 。第三步，使用统计的方法对候选关系实例进行过滤。下面将详细描述这三个步骤。

1) 获取关系模板

获取关系模板分为三个步骤:

- 从语料库中查询获取包含正例的句子: 对于给定的种子 (r, E_1, E_2) ，从语料库中查询同时包含 $E_1 E_2$ 的句子。检索之后会得到一个句子集合 $\{candidate-s_i\}$ ，集合中的每一个句子都会包含 $E_1 E_2$ 。
- 句子过滤: 并非所有通过查询获得的句子所包含的实体都蕴含关系 r ，所以还需要对句子进行过滤。 $E_1 E_2$ 的词距离超过 $minDis$ 的句子过滤^{[28][31]}，例如对于雇佣关系句子“美国总统奥巴马”将被保留，而句子“奥巴马针对基地组织领导人进行的军事打击，究竟在多大程度上削弱了该组织的力量引发了美国民众的激烈讨论”将被过滤。Banko et al.等人^[8]提出一种简单的启发式规则来过滤不包含关系的句子。他们提出在句子中两个实体的依存路径长度不能超过阈值，这种方法取得了很好的效果。通过过滤后，得到句子集合 $\{s_i\}$ 。
- 学习关系模板: Yan 等人^[32]对序列化模板 (Surface Patterns) 和依存模板 (Dependency Patterns) 的训练都有深入的研究。获取关系模板的过程也可以看作是一个学习的过程，让模型自己选择那些模板更能描述关系 r 。

2) 抽取候选关系实例

有了关系模板之后，我们可以训练一个模型 M ，用来在生语料库中抽取关系实

例。我们可以扫描生语料库 D 获取所有包含两个或两个以上实体的句子。然后使用这些模板对所有的句子进行匹配，如果匹配成功则获取一个新的候选关系实例。

3) 过滤关系实例

由于关系种子的数目有限，导致模型有很大噪声。为了减少新的关系实例的错误率，还需要在整个生语料库 D 上进行统计分析，利用生语料库 D 的冗余信息，保留可信度高的候选关系实例。同一个关系实例可能在多个不同的句子中出现，例如雇佣关系的实例（百度，李彦宏）出现在如下句子中：

“百度总裁李彦宏明确表示，百度将在移动互联网方面发力。”

“3月23日消息，据国外媒体报道，百度 CEO 李彦宏表示，该公司计划开发适用于移动设备的操作系统，该战略表明这家领先的中国网络搜索公司将再次追随谷歌。”

这样，可以把出现的句子数目多的关系实例保留下来，当做最终的抽取结果。

1.4.3 开放式实体关系抽取

上述两个研究方向都需要预先确定关系类型体系，然而预先定义一个全面实体关系类型体系是很困难的。开放式实体关系抽取技术^[8]使用实体对上下文中的某些词语来描述实体之间的语义关系，从而避免了构建关系类型体系。开放式实体关系抽取任务是在文本中抽取关系三元组（entity1, relationWords, entity2），其中（entity1, entity2）是存在关系的实体对，relationWords（关系指示词）是上下文中描述实体对的语义关系的词或词序列。在文本“早在去年，腾讯首席执行官马化腾就多次全面阐述了腾讯的发展战略。”中可以抽取关系三元组（腾讯，首席执行官，马化腾）。

英文上的开放式实体关系抽取研究最早在 2007 年被提出来，目前研究方法已经比较成熟。2007 年 Michele Banko 等人^[8]最早提出开放式信息抽取（Open IE, Open Information Extraction）的概念，并且构建了 TextRunner 系统，该系统利用启发式规则从宾州树库中自动构建开放式关系抽取语料，然后利用自动构建的语料训练一个朴素贝叶斯模型识别关系三元组。Fei Wu 等人^[33]提出 WOE 系统使用维基百科中信息框（Infobox）的信息来标注关系抽取语料，基于 WOE 构建语料的方法大大提高了训练语料的质量和数量。Mihai Surdeanu 等人^[34]认为同一个实体对在不同的上下文中呈现出不同的关系，为了解决这个问题提出了 MIML 模型提高自动标注语料的准确率。Anthony Fader 等人^[35]对 TextRunner 系统和 WOE 系统的关系三元组抽取结果进行分析，发现错误的关系三元组主要分为不合逻辑和无意义

两类，为了减少这两类错误，提出了先识别关系指示词的 ReVerb 系统。Limin Yao 等人^[36]认为同一个关系模板可以描述不同的关系，提出了基于 LDA 的关系模板聚类模型，先使用 LDA 模型确定模板的语义类别，再使用层次聚类对这些模板进行聚类，最后形成一个关系类型体系。

在上述介绍的英文的开放式实体关系相关研究中，TextRunner、WOE 以及 ReVerb 都是完整开放式实体关系抽取系统，下面将分别介绍这几个系统。

1) TextRunner

TextRunner 是第一个开放式实体关系抽取系统，不需要人工定义关系类型体系，使用启发式规则在宾州树库中自动标注语料。图 1-2 是 TextRunner 的系统框架图。

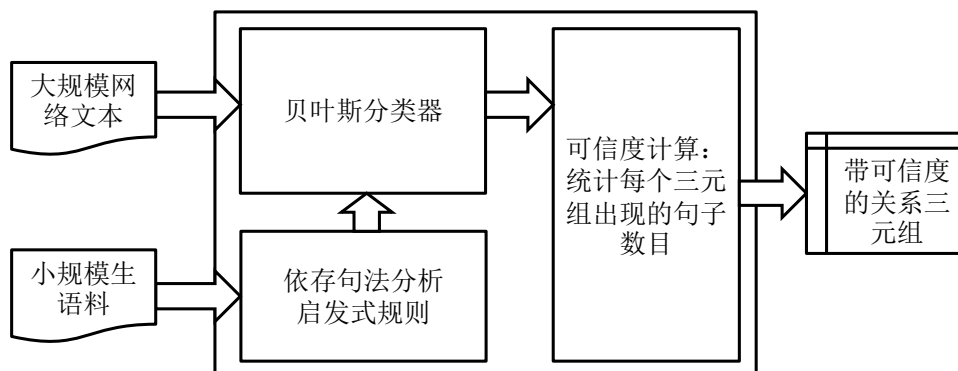


图 1-2 TextRunner 系统框图

Fig. 1-2 Architecture of TextRunner

TextRunner 包含 3 个主要的模块，下面将详细介绍这三个模块。

i. 训练分类器

TextRunner 使用启发式规则自动从宾州树库中构建训练语料，然后再训练 Naïve Bayes 分类器。

a) 利用启发式规则自动构建语料（基于依存句法分析）

- 两个实体的依存距离不能大于 `maxLength`;
- 两个实体的依存路径不能跨越句子边界（例如关系从句是一个句子边界）;
- 两个实体不能是由单独的代词组成;
- 关系指示词是依存两个实体的路径上的动词或动词短语;
- 满足上述前 3 个要求的被标注成正例，否则被标注成反例。

b) 训练 Naïve Bayes 分类器，对每一个三元组(e_i, r_{ij}, e_j)进行分类。其特征如下:

- r_{ij} 的词性标注序列;
- r_{ij} 的长度;
- r_{ij} 中包含的停用词数目;
- 实体是否是专有名词;

- e_i 左边词语的词性;
 - e_j 右边词语的词性。
 - ii. 对大规模 Web 文本进行处理, 对每一个句子中的三元组使用 Naïve Bayes 分类器进行分类, 如果被标注成正例, 那么把三元组存储在数据库中。
 - iii. 计算三元组的可信度
 - a) 合并相似的三元组, 如(e_1 , was developed by, e_2), (e_1 , was originally developed by, e_2)是两个相同的三元组;
 计算合并后三元组在 Web 文本中出现的句子数目, 把句子数目作为三元组的一个可信度。
- 2) WOE

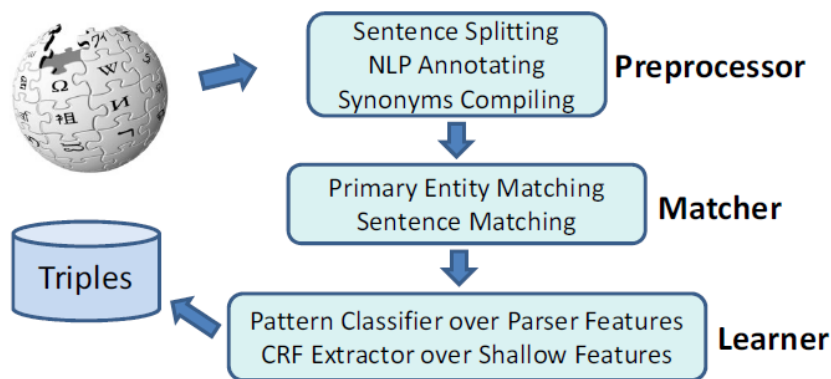


图 1-3 WOE 系统框架图
Fig. 1-3 Architecture of WOE

WOE (Wikipedia-based Open Extractor) 是基于维基百科的一个开放式信息抽取系统, 其关键在于使用维基百科中的信息框自动构建语料库。如图 1-3 所示, WOE 包含三个主要的模块:

- i. 预处理
 - a) 对维基百科中的文本进行分句
 - b) NLP 处理
 - 词性标注
 - 名词短语识别
 - 依存句法分析 (只针对 WOE^{parse})
 - c) 同义词扩展, 找出网页信息框中所有词语的同义词
 - 维基百科中的重定向连接
 - 后向链接。

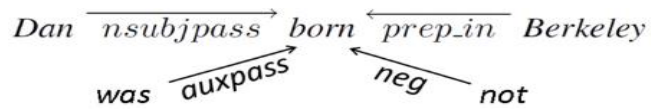
ii. 构建语料库

- a) 对信息框中的每一个（属性，属性值）都在正文中匹配，匹配方法如下：
- 整体匹配，即整个字符串都匹配成功
 - 同义词匹配
 - 前缀匹配，如"Amherst"和"Amherst, Mass"能匹配成功，但是"Mass"和"Amherst, Mass"不能匹配
 - 模板匹配，the <type>。当需要识别 city 的时候，只需要实例化模板 the <city>就可以找到大量的实体
 - 常用的代词
- b) 为了提高语料的质量，需要对 a)中构建的语料进行过滤，过滤规则如下：
- 如果一个属性值被多个句子所匹配，那么把该属性过滤
 - 如果在句子中匹配上的属性值或者文章标题不是名词短语的核心词，那么把这个句子过滤掉
 - 如果句子中匹配上的属性值和文章标题之间跨越了子句，那么把该句子过滤

iii. 训练分类器。WOE 有两个不同的分类器，以满足不同的需求。基于句法特征的分类器是 WOE^{parse} ，基于词性特征的分类器是 WOE^{pos}

 c) WOE^{parse}

- 对句子进行压缩的句法分析，如"Dan was not born in Berkeley"，压缩的句法分析结果是：



- 构建模板集合。依据句法分析结果，用词性代替词，提高模板的召回能力。对"Dan was not born in Berkeley"生成的模板是：

$$N \xrightarrow{nsubjpass} V \xleftarrow{prep} N$$

- 使用公式 $w(p) = \frac{\max(\log(f_p) - \log(f_{min}), 0)}{\log(f_{max}) - \log(f_{min})}$ 对每一个模板进行打分，其中 f_p 代表模板 P 在模板集合中出现的次数。Max 是出现次数最多的模板， f_{min} 是人工设定的一个阈值。

WOE^{pos} 不使用句法，从而速度比 WOE^{parse} 更有优势。其方法是训练一个 CRF 模型，使用 WOE^{parse} 中的模板集合生成正例，如果句子不被模板匹配，那么被标注成反例。

3) ReVerb

和 TextRunner 以及 WOE 这些开放式实体关系抽取系统先识别实体再识别关系词不同, ReVerb 先识别关系词再识别实体。ReVerb 有两个主要的步骤, 如下所述:

- i. 识别关系短语, 使用两个限制:
 - a) 语法限制, 满足下面三条规则的词串作为关系短语

$V \mid VP \mid VW^*P$
$V = \text{verb particle? adv?}$
$W = (\text{noun} \mid \text{adj} \mid \text{adv} \mid \text{pron} \mid \text{det})$
$P = (\text{prep} \mid \text{particle} \mid \text{inf. marker})$

图 1-4 关系指示词的语法限制

Fig. 1-4 Syntactic Constraint for relation words

句子 “Hudson was born in Hampstead, which is a suburb of London” 中找出两个候选短语 “was born in” “is a suburb of”。

b) 词汇限制

- 统计利用 a) 中三条规则找出来的短语在大规模语料中出现的三元组数目;
- 如果出现次数少于阈值, 那么把该短语丢弃。例如 “The Obama administration is offering only modest greenhouse gas reduction targets at the conference.” 中的 “is offering only modest greenhouse gas reduction targets at” 将被该短语过滤掉。

ii. 三元组识别

- a) 找出关系短语左右两边最近的两个名词短语, 构成一个三元组。例如句子 “Hudson was born in Hampstead, which is a suburb of London” 中有两个三元组 (Hudson, was born in, Hampstead) 和 (Hampstead, is a suburb of, London)。
- b) 利用线性加权模型对三元组进行分类。

由于采用的特征是领域无关的, 并且没有词的特征, 所以可以应用到开放域信息抽取中。在训练线性加权模型时, 需要人工标注一部分语料, 在该系统中一共标注了 1000 个句子作为训练语料。

Anthony Fader^[35]对三个系统的性能进行了对比, 如图 1-5 所示。

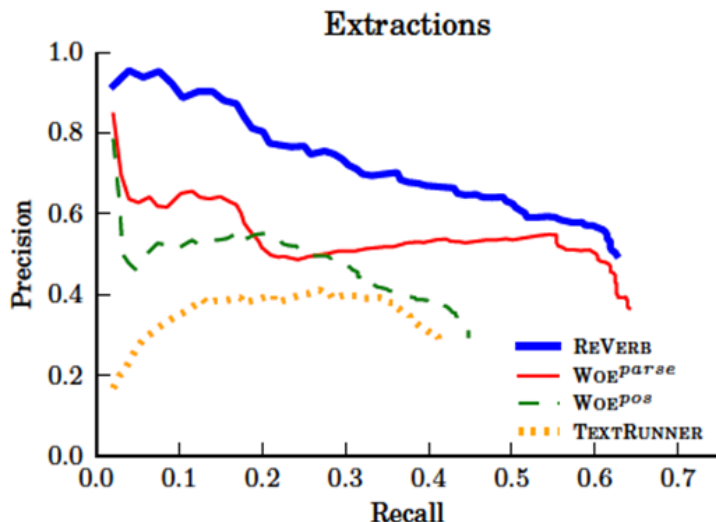


图 1-5 三个系统的 PR 曲线

Fig. 1-5 PR curve of three systems

中文的开放式实体关系抽取相关研究还比较少,中文的语言现象和英文的语言现象相差比较多,无法直接把英文上的开放式实体关系抽取方法直接移植到中文上来。王莉峰^[37]提出了领域自适应的中文实体关系抽取方法,结合半指导和无指导的学习方法解决关系类型自动发现、关系种子集自动构建、关系描述模式挖掘和关系元组抽取等问题,并且应用到音乐领域人与人之间的关系三元组识别任务上,取得了不错的效果。

1.5 本文的主要研究内容

本文将探索开放式中文实体关系抽取的解决方案,针对不同的应用场景分别提出有指导和无指导的方法来解决关系三元组抽取问题,并且运用聚类技术解决开放式实体关系类型体系自动构建的问题。

本文的研究工作流程如图 1-6 所示。具体地,本文各章节安排如下:

第 1 章,首先介绍本研究课题的来源和研究背景,接着探讨研究的目的和意义,然后详细介绍实体关系抽取的研究方向以及方法,并且分析这些方法的不足,在此基础上,提出本文的主要研究内容。

第 2 章,针对句子级的开放式关系抽取问题,提出基于有指导的方法,同时制定了中文开放式实体关系语料的标注规范,标注开放式实体关系抽取语料。在对语料细致分析的基础上,提出两种有指导的关系三元组抽取方法。

第 3 章,针对互联网的开放式实体关系抽取问题,提出面向互联网的无指导开放式中文实体关系抽取(UnCORE: Unsupervised Chinese Open Entity Relation

Extraction for the Web) 方法, 首先使用实体之间的距离限制和关系指示词的位置限制获取候选关系三元组, 然后采用全局排序和类型排序的方法来挖掘关系指示词, 最后使用关系指示词和句式规则对关系三元组进行过滤。在不同的领域使用 UnCORE 方法, 以证实方法的鲁棒性。

第4章, 在前两章的工作基础上, 研究和探讨了基于关系指示词聚类的开放式实体关系类型体系构建的方法。我们以 PER-PER 的关系指示词集合为处理对象, 尝试了不同的相似度计算算法和关系指示词聚类算法, 最终形成一个类型丰富的关系类型体系。

第5章, 为了把本文的相关研究成果展示给研究人员, 在第二章的基础上搭建面向句子的开放式实体关系抽取系统, 系统从输入的自然语言文本中抽取关系三元组, 并且把抽取结果以网页的形式展现; 同时, 对第三章从互联网文本中挖掘的大量关系三元组构建索引, 搭建了开放式实体关系三元组查询系统, 以清晰直观的方式展示关系三元组。

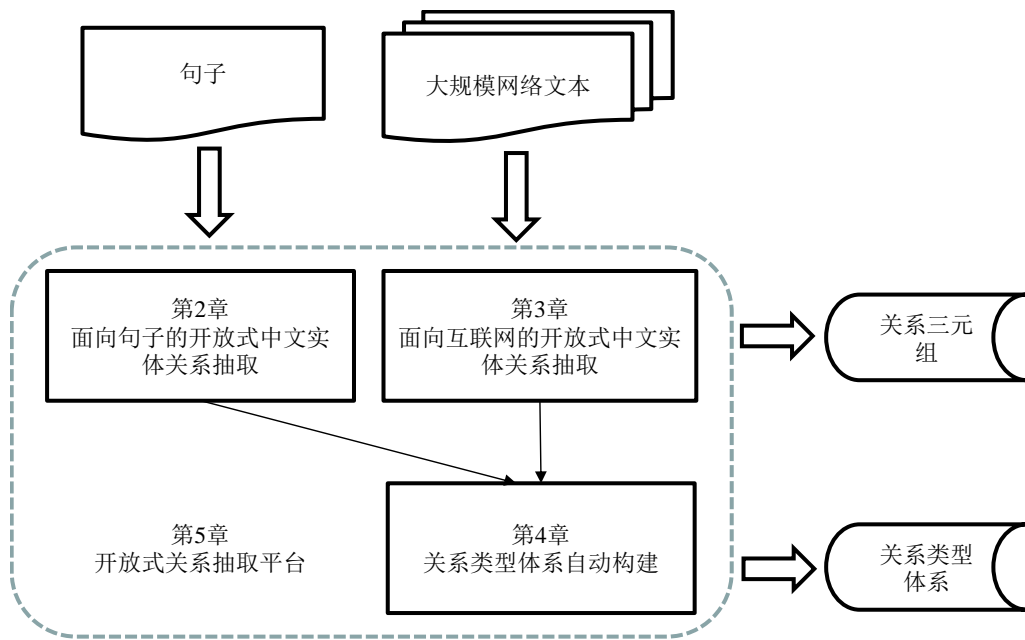


图 1-6 本文研究工作流程
Fig. 1-6 Architecture of this paper

第2章 面向句子的开放式中文实体关系抽取

2.1 引言

基于机器学习的方法在传统的关系抽取上取得了很好的效果，那么是否可以使用机器学习算法来解决开放式的实体关系抽取任务呢？答案是肯定的。但是不同于传统的实体关系抽取任务，开放式实体关系抽取任务是在文本中抽取关系三元组（实体 1，关系指示词，实体 2），需要识别出关系指示词。所以把传统的关系抽取任务看作一个分类问题的做法在开放式实体关系抽取任务中不再适用，在本章中，我们把识别关系指示词的问题看作是序列化标注的问题。

2.2 语料构建

有指导的方法需要带开放式实体关系标记的语料库，但是目前还没有中文的开放式实体关系语料，所以我们构建了一个开放式实体关系抽取语料库。首先从 Ontonotes 4.0 上选取了 1000 篇文档，大部分文档已经带有命名实体标记，我们人工对没有标注命名实体标记的文档进行标注。

在标注语料过程中，首先要指出这次标注的内容。在标注开放式实体关系的时候需要标注的内容有 5 个：

- 1) 关系元素 1：构成关系的第一个实体。
- 2) 关系元素 2：构成关系的第二个实体。
- 3) ACE 实体关系类型：如果实体之间构成的关系类型能用 ACE 实体关系类型来描述，那么标注 ACE 实体关系类型。
- 4) 关系指示词：在上下文中能描述两个实体构成关系的词语，可以是多个词语。
- 5) 关系是否对称：两个关系元素的相对位置改变之后，是否还构成相同的关系，如果构成相同关系那么就对错，否则不对称。

其中关系元素 1、关系元素 2、关系指示词是句子中的词片段，“ACE 实体关系类型”和“关系是否对称”为标注者添加的信息。所以对每一个开放式关系实例，可以用一个五元组来描述，五元组形式为 `<arg1,arg2, ACEType, relationWord, isSymmetrical>`。

本次标注的输入是一个已经分词、命名实体识别的词序列，输出是序列中所有的关系五元组。

输入:微软公司/ORG 董事长 比尔·盖茨/PER 出生 于 1955 年 10 月/TIME 。

输出: <微软公司, 比尔·盖茨, 机构关系, 董事长, 否>

<比尔·盖茨, 1955 年 10 月, 其他, 出生, 否>

表 2-1 是一些实体关系的标注样例, 指示词中用引号引起来的词语为标注者添加的指示词。

表 2-1 开放式实体关系标注实例
Table 2-1 Samples of open relation instance

例句	关系元素 1	关系元素 2	ACE 关系类型	指示词	对称性
华西金塔矗立在苏南平原	华西金塔	苏南平原	物理 处于	矗立	否
他曾经效力于中央电视台	他	中央电视台	机构关系 雇佣	曾经效力	否
他曾经住在巴黎	他	巴黎	一般关系 市民	曾经住在	否
姚明是易建联的朋友	姚明	易建联	持续的-私人的	朋友	是
毛泽东出生于 1893 年	毛泽东	1893 年	其他	出生	否
《花蝴蝶》中蔡依林为了让影子呈现蝴蝶展翅高飞的效果, 坚持吊威亚近一小时后, 她的右腰出现了深色瘀青	蔡依林	花蝴蝶	其他	“歌曲”	否

为了标注的规范性, 我们对标注过程中遇到的一些问题进行分类, 并且制定了标注规范, 如下:

1) 实体关系

实体之间存在特定的语义关系的都要标注出来。例如“朗讯是阿尔卡特的子公司”中存在实体关系实例<阿尔卡特, 朗讯, 整体部分, 子公司, 否>。

2) 关系元素

关系元素 1 和关系元素 2 是构成特定关系的两个关系元素。关系元素包括命名实体和代词。

3) 关系指示词的类型以及标注方法

关系指示词是句子中用来出发两个实体（或者代词）之间的词序列, 有时在句

子中找不出用来指示两个实体的关系词语，需要添加一个合适的词语来指示该关系。关系指示词为名词、动词以及短语。

- 动词：“毛泽东出生于1893年”中的“出生”指示“毛泽东”和“1893年”两个实体之间存在“出生日期”的关系；
- 名词：“林俊杰携手师妹金莎一起来到上海”中的“师妹”指示“林俊杰”和“金莎”存在“同门师兄妹”的关系；
- 短语：“刘欢和莎拉·布莱曼共同演绎了主题歌《我和你》”中的“共同演绎”指示“刘欢”和“莎拉·布莱曼”存在“合作关系”。
- 无关系指示词：“《花蝴蝶》中蔡依林为了让影子呈现蝴蝶展翅高飞的效果，坚持吊威亚近一小时后，她的右腰出现了深色瘀青”中的“花蝴蝶”和“蔡依林”存在实体关系，但是在句子中找不出指示实体间关系的词语，需要标注者添加一个合适的词语（最好是名词），例如上例中可以添加“歌曲”来指示该关系。

4) ACE 实体关系类型以及关系指示词与 ACE 实体类型的映射

句子中指示词(relationWord)的可以归类到{人工制品，一般关系，机构关系，部分-整体，人-社会，物理，其他}这几种关系类型中。下面是对这几个关系类型的具体描述。

- 人工制品：使用者-拥有者，发明者-制造者；
- 一般关系：市民-居民，宗教团体-种族、机构所在地；
- 机构关系：雇佣、创建者、所有者、学生-校友、运动-团体、投资者-股东、会员；
- 部分整体：人工制品、地理、子公司；
- 人-社会：家庭、持续的-私人的；
- 物理：处于，临近；
- 其他：其他。

对其中一些关系的说明：

- 物理
 - 处于：江、河、湖、山等的位置，arg1 位于 arg2，区别于整体部分关系；
 - ◆ 洞庭湖位于湖南。
 - ◆ 他来到哈尔滨。
 - ◆ “哈尔滨位于黑龙江”属于“整体部分”中的“地理关系”，不在本关系中。
 - 临近：位置不是相邻的关系；

- ◆ 哈尔滨工业大学在“脑汇附近。”
- 人-社会
 - 家庭：亲属关系。
 - ◆ 刘德华的老婆是朱丽倩。
 - 持续的-私人的：人和人之间非亲属关系
 - ◆ 姚明是易建联的朋友。
 - ◆ 他是我的上级。
- 机构关系：ORG-PER
 - 雇佣
 - ◆ 他是 NEC 的员工。
 - 创建者
 - ◆ 马化腾创建了腾讯公司。
 - 所有者
 - ◆ 他继承了父亲的所有财产，从而拥有 ABC 公司。
 - 学生-校友
 - ◆ 1988 年，李开复获卡内基梅隆大学计算机学博士学位。
 - 运动-团体
 - ◆ 来自中国队的林丹获得了羽毛球冠军。
 - 投资者-股东
 - ◆ 巴菲特从 2006 年开始投资康菲石油。
 - 会员
 - ◆ 他购买了腾讯公司的黄钻会员。
- 人工制品：人工制品为 arg1
 - 使用者-拥有者：
 - ◆ 我的 iphone 4S。
 - 发明者-制造者
 - ◆ 统计自然语言处理基础是 Manning 写的。
 - ◆ 苹果公司制造了 iphone 4s。
- 一般关系
 - 市民-居民：PER-LOC
 - ◆ 他是中国人。
 - ◆ 他出生于哈尔滨。
 - 宗教团体-种族：PER-NC

- ◆ 小明是汉族人。
- ◆ 他信仰基督教。
- 机构所在地：ORG-LOC
- 部分-整体：整体为 arg1，部分为 arg2
 - 人工制品：人工制品的整体部分
 - 地理：地理上的整体部分，LOC-LOC
 - ◆ 哈尔滨位于黑龙江。
 - ◆ 黄岩岛是中国的领土。
- 子公司：ORG-ORG，母公司在前，子公司在后。

5) 对称性

关系元素 1 和关系元素 2 可以交换位置的关系实例称为具有对称性的关系元组。例如“姚明是易建联的朋友”中的朋友关系具有对称性而“叶惠美是周杰伦的母亲。”中的“母亲”关系不具有对称性。

如果关系具有对称性，那么在句子中先出现的实体被标注为 arg1，后出现的实体被标注为 arg2。如果关系不具有对称性，可以用以下两个规则来确定 arg1 和 arg2。

- 可以用所有格的方法来确定，即“<arg1>的<relationWord>是<arg2>”；
- 考察“<arg1><relationWord><arg2>”所构成的句子是否通顺。

2.3 有指导的开放式实体关系抽取

在抽关系取三元组（实体 1，关系指示词，实体 2，）的时候，有两种不同元素需要从原本中抽取出来，一个是实体对（实体 1，实体 2），还有关系指示词。因此我们设置了两种不同的方法来识别关系三元组：先识别实体对的方法和先识别关系指示词的方法。

2.3.1 语料分析

为了更好的关系三元组抽取构建模型，我们对语料进行了分析，统计了语料中关系三元组的关系指示词和实体对的相对位置信息。表 2-2 显示了关系指示词在句子中的位置。关系指示词的位置分为以下几种：

- 1) 两个实体之间：句子中“<ORG>哈尔滨工业大学</ORG><RELATION_WORD>校长</RELATION_WORD><PER>王树国</PER>来到计算机学院。”的关系指

- 示词“校长”在实体“哈尔滨工业大学”和实体“王树国”中间；
- 2) 实体右边：句子“<PER>梁朝伟</PER>是<PER>刘嘉玲</PER>的<RELATION_WORD>老公<RELATION_WORD>。”的关系指示词“老公”在两个实体的右边；
 - 3) 实体左边：句子“作为<RELATION_WORD>董事长<RELATION_WORD>，<PER>李彦宏</PER>在<ORG>百度</ORG>有别人不可替代的作用。”的指示词“董事长”在两个实体的左边；
 - 4) 没有指示词：句子“<ORG>中国</ORG>的<PER>刘翔</PER>在田径比赛中取得了可喜的成绩。”中没有关系指示词，但是两个实体确实存在关系（刘翔，国籍，中国）；
 - 5) 错误：不在上述四种类型当中的，例如“<ORG>铁道<RELATION_WORD>部长</RELATION_WORD></ORG> </PER>傅志寰</PER>。”中“铁道部长”是实体，但是关系指示词是“部长”，两者有重合，这是语料标注错误。

表 2-2 指示词在句子中的位置分布

Table 2-2 The distribution of relation triples with the position of relation word

关系指示词的位置	实例数目（个）	比例（%）
两个实体之间	3177	75.36
实体右边	609	14.44
实体左边	160	3.80
没有指示词	240	5.69
错误	30	0.71

从表 2-2 中可以看出，绝大部分（93.6%）存在关系的实例在句子中都能找到一个关系指示词来标识实体之间的关系，这也验证了用三元组来描述一个关系实例的是可行的。75.36%的关系实例的关系指示词在两个实体的中间，14.44%的关系实例的关系指示词在实体的右边，这两者占总关系实例数目的 89.80%，覆盖了大部分（95.94%）存在关系指示词的关系实例。基于上面的发现，我们在构建关系抽取模型的时候，只需考虑“关系指示词在两个实体之间”和“关系指示词在两个实体右边”这两种情况即可，在保证关系三元组抽取效果的同时也简化了模型的复杂程度。

为了更加精准的找到关系指示词，我们统计了“关系指示词在实体右边”的时候，关系指示词与实体 2 距离取不同的值时关系三元组的数目，如表 2-3 所示。可以发现右边的前三个词可以覆盖大部分（80.92%）情况，所以在实验的时候，只考虑实体 2 右边三个词。

表 2-3 关系三元组的数目和指示与实体 2 距离的关系
Table 2-3 The distribution of relation triples with the distance of the sencond entity

距离	0	1	2	3	4	5	6	7	8	大于 8
数目	292	145	55	35	17	17	19	11	6	11

2.3.2 先识别实体对的方法

在图 2-1 中，实线部分是系统的训练过程，虚线部分是预测过程。训练数据和测试文本都先通过特征提取器提取特征。提取特征之后，先通过实体对识别模型，判断实体之间是否存在关系；如果实体之间存在关系，则再通过关系指示词识别模型把描述实体之间关系的词语标注出来。

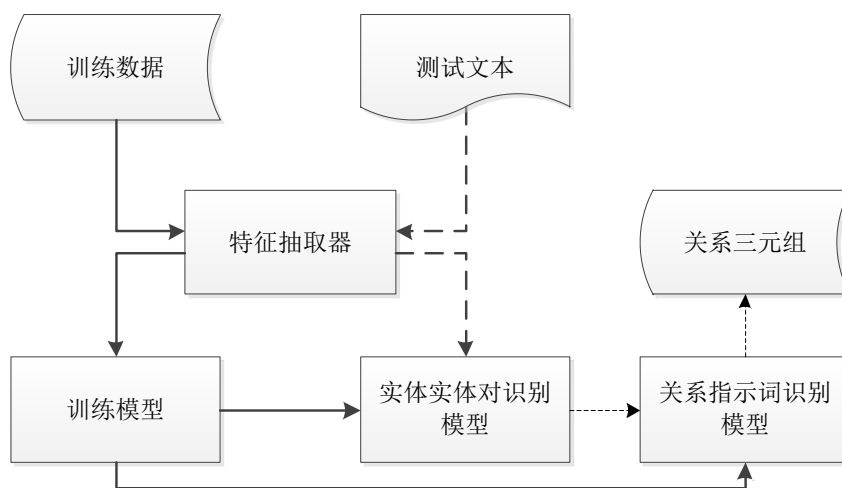


图 2-1 先识别实体对的算法框架图

Fig. 2-1 Architecture of the first method

1) 实体对识别模型

实体对识别模型可以判断句子中两个实体是否存在关系，使用最大熵算法训练。包含以下一些特征：

- 实体本身特征：实体 1 的类型，实体 2 的类型，实体 1 的词序列，实体 2 的词序列；
- 两个实体之间的特征：是否有其他的实体，是否相邻，两个实体之间的每个词语；
- 实体两边的特征：实体 1 左边 3 个词和词性，实体 2 右边 3 个词和词性。

在使用实体本身的特征时，本文使用了实体的词序列，而没有使用实体本身。这是由于实体分成若干词语之后，特征的泛化能力更强。假如训练语料中包含实体“哈尔滨工业大学”，其词序列的特征是{哈尔滨，工业，大学}，在测试过程中

遇到实体“哈尔滨工程大学”，其用词序列的特征为{哈尔滨，工程，大学}，可以看出，两个实体词序列的特征有一部分是相同的。

2) 关系指示词识别模型

当 1) 判断出两个实体存在关系时，那么就需要把指示他们关系的词语抽取出来。由于关系指示词语可以使一个词序列，所以理所当然地把它看作一个序列化标注的问题。而序列化标注使用最多的模型是 CRF，所以本文训练了一个 CRF 模型抽取特征词。其训练语料是只包括正例（即只有存在关系的实体对），没有反例。

在 2.3.1 小节中，我们提到过，大部分关系三元组的关系指示词存在于实体之间和实体右边 2 个词，所以 CRF 标注关系指示词的候选限定于这个范围之内。CRF 模型使用的特征有：词、词性、词和词性的组合、是否是实体。这些都是 unigram 的特征，没有使用 bigram 的特征。

我们使用{B, I, O, E, S}三类标签标注候选关系指示词，B 代表这个词语是关系指示词序列的开始，I 代表这个词语在关系指示词的中间位置，O 代表这个词语不是关系指示词的一部分，E 代表这个词语是关系指示词的结束，S 代表这个词语是一个完整的关系指示词。下面是一个标注样例：

“美国/O 代表团/B 的/I 团长/E 希尔/O 在/O 呢/O ， /O”

其中，（美国，希尔）是存在关系的实体对，“代表团的团长”是关系指示词。

2.3.3 先识别关系指示词的方法

本方法与 2.3.2 所描述的方法最大不同是，本方法不在需要判断实体之间是否存在关系。

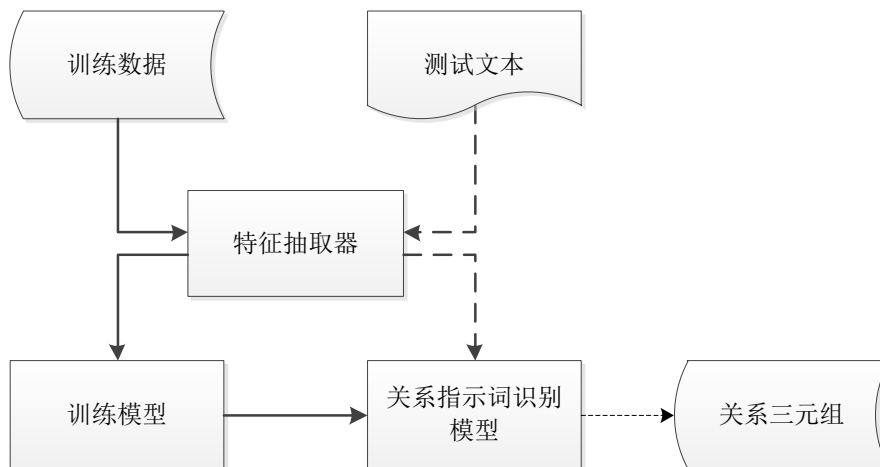


图 2-2 先识别关系指示词的算法框架图

Fig. 2-2 Architecture of the first method

如图 2-2 所示，对句子中任意两个实体，先抽取特征，然后使用识别关系指示

词，如果句子中有词语被标注成关系指示词，那么这两个实体就构成关系，否则实体之间不存在关系。图中实线代表训练过程，虚线代表预测过程。

关系指示词识别模型使用的特征有：词、词性、词和词性的组合、词语是否是实体。

2.4 实验结果及其分析

训练和测试语料是我们自己标注的信息抽取语料，语料构建规范如 2.2 节所述。标注信息包括分词、命名实体、名词复合短语、实体共指、实体关系、事件以及篇章语义 7 种，在本章试验中使用的信息有分词、命名实体、实体共指和实体关系 4 种。

实验使用了其中的 805 篇标注文档。训练样例是以实体对为单位的，其中正例 3658 个、反例 95401 个。这样正反例严重不平衡。所以使用了两条规则对反例进行过滤。

- 1) 如果两个实体存在共指，那么这两个实体不存在关系
- 2) 对任意 $i > j$ 并且实体 i 和实体 j 共指，如果 $k > j$ ，则实体 k 和实体 i 不存在关系。

对语料使用规则 1) 之后，剩下正例 3658 个，反例 91285 个。同时使用规则 1) 和 2) 之后，剩下正例 3656 个，反例 86323 个。可以看出，使用规则后，反例有所减少，但还是存在数据不平衡的问题。

2.4.1 先识别实体对的方法

先识别实体对的方法包含两个模型，我们对每一个模型都有评价。

表 2-4 判断两个实体之间是否存在关系的分类效果

Table 2-4 The performance of maxent model

类别	准确率 (%)	召回率 (%)	F 值 (%)
实体之间存在关系	62	21	31
实体之间不存在关系	96	99	98

从表 2-4 可以看出，数据不平衡问题在对分类结果又很大的结果。数量占优的一方无论是准确率还是召回率上都完全超过占劣势的一方。而所关注的恰恰是数量较少的一方（存在关系的三元组），其准确率只有 62%，这给后续实验带来的累积错误太大。

表 2-5 是在两个实体存在关系的前提下关系指示词的抽取效果，评价设置了三种不同标准。

表 2-5 关系指示词抽取实验结果

Table 2-5 The performance of relation words extraction

评测标准	准确率 (%)	召回率 (%)	F 值 (%)
精确匹配	73.14	70.23	71.66
模糊匹配 1	74.67	71.70	73.16
模糊匹配 2	75.98	72.96	74.44

- 1) 精确匹配: 模型的抽取结果要与标准结果完全一致。
- 2) 模糊匹配 1: CRF 抽取出来的特征词是标准答案的一部分, 如表 2-6 所示。

表 2-6 模糊匹配 1

Table 2-6 Fuzzy matching 1

词	标准答案	预测结果
美国	0	0
代表团	B	0
的	I	0
团长	E	S
希尔	0	0
在	0	0
呃	0	0
,	0	0

- 3) 模糊匹配 2: 标准答案的特征词是 CRF 抽取的特征词的一部分, 如表 2-7 所示。

表 2-7 模糊匹配 2

Table 2-7 Fuzzy matching 2

词	标准答案	预测结果
悉尼	0	0
奥运会	0	B
女子	B	I
体操	I	I
金牌	I	I
获得者	E	E
刘璇	0	0
告诉	0	0
记者	0	0
,	0	0

表 2-5 所展示的是判断实体之间存在关系全都正确前提下的关系指示词识别结果。如果在使用模型先判断两个实体是否构成关系，然后再识别有关系的实体对的关系指示词，将会累积两个模型的错误。级联后的关系三元组抽取结果评价如表 2-8 所示，可以看出，无论准确率和召回率的值都很低，这可能是复杂的模型会带来过拟合的现象。

表 2-8 先识别实体方法的关系三元组抽取实验结果

Table 2-8 The performance of second method

评测标准	准确率 (%)	召回率 (%)	F 值 (%)
精确匹配	45.35	14.75	22.26
模糊匹配 1	46.30	15.06	22.72
模糊匹配 2	47.11	15.32	23.12

2.4.2 先识别关系指示词的方法

表 2-9 是先识别关系指示词方法的关系三元组抽取实验结果，对比表 2-8 可知，先识别关系指示词的方法比先识别实体对的方法效果在 F 值上要高 47% 以上，其提高幅度是很明显的。其原因是先识别关系指示词的方法只是用了 CRF 模型，在训练时同时使用了正反例信息，比先识别实体对方法中关系指示词识别模块的信息多。

表 2-9 先识别关系指示词方法的三元组抽取实验结果

Table 2-9 The performance of second method

评测标准	准确率 (%)	召回率 (%)	F 值 (%)
精确匹配	87.80	45.28	59.75
模糊匹配 1	87.80	45.28	59.75
模糊匹配 2	90.24	46.54	61.41

2.4.3 实验结果分析

表 2-10 两个方法的实验结果对比

Table 2-10 Compare of two methods

方法	准确率 (%)	召回率 (%)	F 值 (%)
先识别实体对	47.10	15.32	23.12
先识别关系指示词	90.24	46.54	61.41

表 2-10 是本章两个方法的实验结果对比，可以看出先识别关系指示词的方法

的关系三元组抽取实验结果比先识别实体对的关系三元组抽取实验结果的 F 值提高了 38.29%，这是由于先识别实体对的方法中使用了最大熵模型，数据不平衡问题严重影响了最大熵模型的效果。而先识别关系指示词的方法中只是用了 CRF 模型，数据不平衡的问题对 CRF 模型影响没有造成很大的影响。

2.5 本章小结

本章构建了一个开放式实体关系语料库，制定了语料标注规范，得到一个标准的评价集。通过分析开放式实体关系语料，证实了使用三元组来描述一个关系实例是可行的。

我们把开放式实体关系抽取任务分成两个子问题：实体对识别和关系指示词识别。针对两个子问题的解决先后顺序，分别设计了两种不同的解决方案：先识别实体对的方案和先识别关系指示词的方案。由于采用了有指导的方法，为了增强模型的移植能力，我们设计了泛化能力较强的特征：使用词性、实体的词序列等特征。对两种不同的方案进行实验，发现先识别关系指示词方法的三元组抽取结果的 F 值达到 61.41%，高于先识别实体对的方法，同时还对原因进行了分析。综上所述，使用有指导的方法在有效的解决句子级的开放式实体关系抽取任务的同时，还避免了传统关系抽取中预先定义关系类型体系的问题。

第3章 面向互联网的开放式中文实体关系抽取

3.1 引言

有指导的方法需要构建语料库，当从一个领域移植到另外一个领域时需要重新标注语料，消耗大量的人力资源。同时机器学习算法时间复杂度高，无法适用于处理大规模的网络文本。为了适应快速增长的网络文本，本章提出面向网络文本的无指导的开放式中文实体关系抽取方法。

我们通过分析中文语料库，发现同一个关系指示词往往只出现在特定的实体对类型的三元组中，例如“首席执行官”出现在实体对类型为（机构名，人名）的三元组中，“爸爸”出现在实体对类型为（人名，人名）的关系三元组中。基于上述发现，本章提出一种新颖的面向互联网的无指导开放式实体关系抽取(UnCORE: Unsupervised Chinese Open Entity Relation Extraction for the Web)方法，主要研究人、机构、地点之间的实体关系开放式描述。UnCORE 首先使用实体之间的距离限制和关系指示词的位置限制获取候选关系三元组，然后采用全局排序和类型排序的方法来挖掘关系指示词，最后使用关系指示词和句式规则对关系三元组进行过滤。

3.2 无指导的开放式实体关系抽取

UnCORE 的核心思想是从大规模网络文本中通过启发式规则获取候选三元组，然后从候选三元组中自动挖掘关系指示词，最后利用关系指示词和句式规则过滤三元组。

如图 3-1 所示，UnCORE 的输入是大规模网页、输出是从网页文本中抽取的关系三元组。UnCORE 共包含 4 个模块：

- 1) 预处理模块
- 2) 生成候选三元组模块
- 3) 生成关系指示词词表模块
- 4) 后处理模块

预处理模型对网页进行正文提取，然后把正文转换成带自然语言标记（断句、分词、词性标注、命名实体识别）的句子集合。生成候选三元组模块在句子集合中使用两类限制条件获取候选三元组。生成关系指示词词表模块使用全局排序和

领域排序的方法在候选三元组集合中挖掘关系指示词词表。后处理模块是对候选三元组集合进行过滤和完善，最终得到大规模的关系三元组。

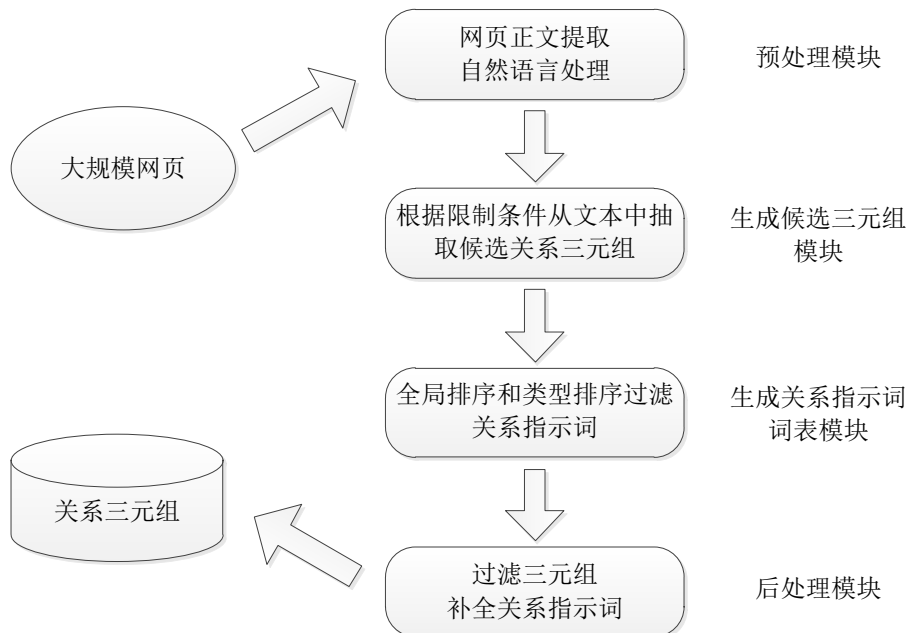


图 3-1 面向互联网的开放式中文实体关系抽取模型

Fig. 3-1 architecture of UnCORE

3.2.1 预处理

预处理模块从网页中获取正文信息并转换成带有自然语言处理标记的句子集合，包含网页正文提取和自然语言处理两个步骤，下面我们将分别介绍这两个步骤。

1) 网页正文提取

使用基于文本行分布的正文抽取¹方法对网页进行处理，抽取出网页中的正文文本。在网页正文提取结果中，随机选取 300 个百科网页上进行人工判断，基于文本行分布的正文抽取方法准确率达到 95% 以上。

2) 自然语言处理

使用哈尔滨工业大学社会计算与信息检索研究中心发布的语言技术平台^[38]（LTP, Language Technology Platform）对网页正文进行断句、分词、词性标注和命名实体识别。

对网络文本进行命名实体识别的时候，发现很多机构名都不能识别出来，这是由于在 LTP 集成的命名实体识别模型是使用人民日报语料训练的，导致在处理网

¹ <https://code.google.com/p/cx-extractor/>

络文本时的机构名召回率太低。为了提高实体的召回率，我们通过百度百科²构建了一个机构名列表，构建词表的核心思想是百度百科中的每个词条都有一个开放的标签集合，如果标签集合中出现“公司”、“学校”等类似标签，那么就认为这个词条是一个机构。从百度百科中抽取的机构名列表共包含 19286 个机构名。使用机构名列表的规则很简单：如果文本中的某个词语在机构名列表中，那么就认为这个词是机构名。这样可以召回大量的机构名，提高命名实体的召回率。

3.2.2 生成候选三元组

为了更好的刻画关系三元组抽取模型，同时也为了提高候选三元组抽取的准确率，我们对开放式实体关系语料进行了更细致的分析。通过分析语料，我们提出了两个生成候选关系三元组的限制条件：实体之间的距离限制和关系指示词的位置限制。

1) 实体之间的距离限制

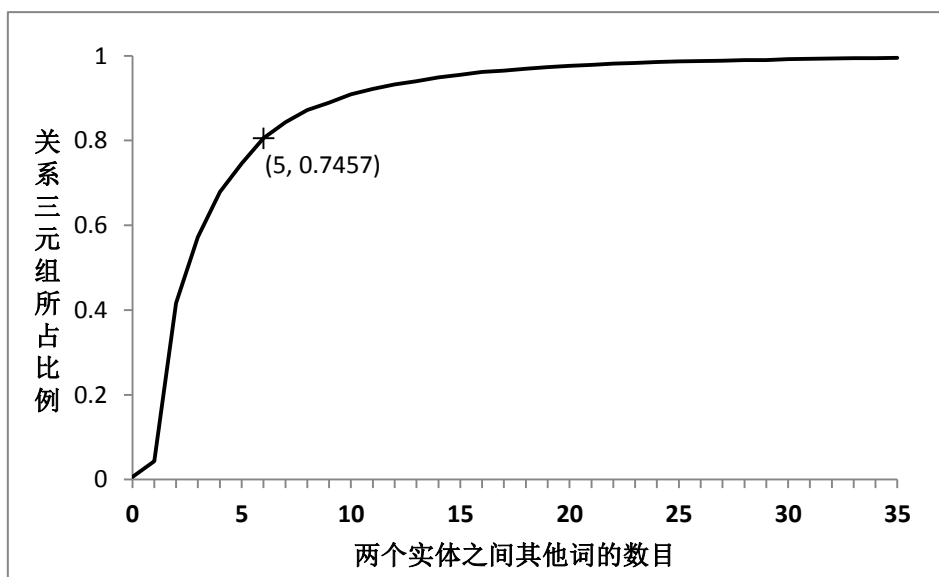


图 3-2 关系三元组数目在词距离上的分布情况

Fig. 3-2 The distribution of relation triples with different word distance

图 3-2 上点 (5, 0.7457) 表示两个实体之间词数目小于等于 5 的关系三元组数目占总的关系三元组数目的 74.57%。从图 3-2 可以看出，当词的数目小于某个值的时候，关系三元组的数量随着词距离增大而急剧上升；而当词的数目超过这个值的时候，随着词的数目的增多关系三元组数量增加幅度越来越小。也就是说大部分存在关系的实体对之间的词距离很小。因此，在生成候选关系三元组的时候

² <http://baike.baidu.com/>

候，我们规定候选三元组的两个实体之间词的数目不能超过 maxDistance 。

图 3-3 上点 (4, 0.9855) 表示两个实体之间其他实体数目小于等于 4 的关系实例数目占总关系三元组数目的 98.55%。从图 3-3 中的曲线可以看出，关系实例的增长速度一直比较缓慢，也说明了实体之间其他实体数量越少越有可能存在关系，所以，在生成候选关系三元组时，本文规定实体之间其他实体数量不能超过 maxEntityDistance 。

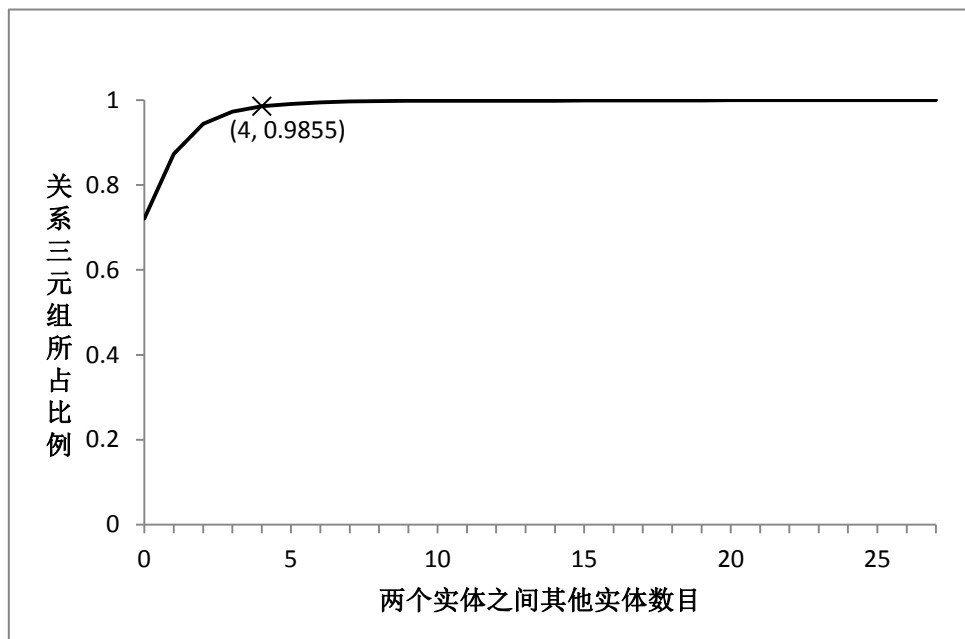


图 3-3 关系三元组数目在实体距离上的分布

Fig. 3-3 The distribution of relation triples with different entity distance

2) 关系指示词的位置限制

在 2.3.1 节中我们提到过，75.36% 的关系实例的关系指示词在两个实体的中间，这说明两个实体中间的词语很有可能是关系指示词。当指示词在实体 1 的左边或者实体 2 的右边时，关系指示词靠近实体的情况占绝大部分的关系三元组。同时我们对关系指示词的词性进行统计，当单个词语作为关系指示词的时候，该词语的词性往往是动词和名词。基于上述发现，我们制定了以下限制条件来抽取候选关系指示词：

- 实体之间的名词和动词
- 第一个实体左边 leftWordNumber 个名词和动词
- 第二个实体右边 rightWordNumber 个名词和动词

通过关系指示词的位置限制，在保证三元组抽取数量的同时，还提高了候选三元组的准确率。

3.2.3 生成关系指示词词表

通过生成候选关系三元组，可以得到候选关系指示词集合。但是候选关系指示词集合中包含了大量的噪声，为了提高关系指示词抽取的准确率，我们对候选关系指示词集合进行了排序和过滤，并且针对每个实体对类型生成一个关系指示词词表。

1) 全局关系指示词排序

前文已经指出同一个关系指示词往往只出现在特定实体对类型的关系三元组中，换一种说法就是关系指示词可以区分不同的实体对类型的关系三元组，区分能力越强的词语越可能是关系指示词。信息增益值可以评价词语的区分能力，信息增益的计算公式(3-1)所示。

$$IG(p) = H(types) - H(types|rel) \quad (3-1)$$

$$H(types) = - \sum_{t \in types} p(t) \log p(t) \quad (3-2)$$

$$H(types|p) = - \sum_{t \in Types} p(rel, t) \log p(t|rel) - \sum_{t \in Types} p(\overline{rel}, t) \log p(t|\overline{rel}) \quad (3-3)$$

其中 rel 表示候选关系指示词， t 表示实体对类型， $t \in types$ 。统计发现，与人相关的关系类型比较丰富，所以本文只关注 $types = \{PER-PER, PER-ORG, PER-LOC, ORG-PER, LOC-PER\}$ 。

使用公式(3-1)可以对关系指示词进行全局排序，其排名靠前的词语可能是关系指示词。

2) 类型关系指示词排序

信息增益能找到指示实体关系的词语，但是不能说明该词语是指示哪一类实体对类型的关系，所以必须使用类型（实体对类型）打分公式来评价一个词语是否能描述特定实体对类型的关系。公式(3-4)计算的是关系指示词 rel 描述实体对类型 t 的实体关系的能力。

$$score(rel, t) = p(t|rel) \log c(rel, t) \quad (3-4)$$

公式(3-4)中 $p(rel|t)$ 保证了指示词 rel 要在实体对类型 t 上出现概率高，才能使 $score(rel, t)$ 的值大；而 $\log c(rel, t)$ 要求 rel 和实体对类型 t 共现次数多，才能使 $score(rel, t)$ 的值大。

具体地，“总裁”在实体对 ORG-PER 中出现的概率比在其他实体对类型中出现的概率值大，并且“总裁”在实体对 ORG-PER 中出现的频率很高，所以使用公式(3-5)计算 $\text{score}(\text{总裁}|\text{ORG-PER})$ 的值很大。

3) 过滤关系指示词

基于全局关系指示词排序和类型关系指示词排序的方法，可以对关系指示词进行过滤，最终生成每个实体对类型的关系指示词词表。

生成关系指示词词表的算法如下：

算法 3-1：生成关系指示词词表

输入：候选关系指示词集合 $\text{CandidateRelationWords}$ ， $\text{IG}(\text{rel})$ ， $\text{score}(\text{rel}, t)$ ， types

输出：关系指示词词表 $\{\text{RelationWords}(t)|t \in \text{types}\}$

步骤：

1. 令集合 $\text{IGCandidateRelationWords}$ 为 $\text{CandidateRelationWords}$ 按照 $\text{IG}(\text{rel})$ 值降序排序结果
 2. 令集合 IGList 为 $\text{IGCandidateRelationWords}$ 的前 N 个元素
 3. 对集合 types 中的每个元素 t
 - 3.1. 令集合 $\text{scoreCandidateRelationWords}(t)$ 为 $\text{CandidateRelationWords}$ 按照 $\text{score}(\text{rel}, t)$ 值降序排序结果
 - 3.2. 令集合 $\text{scoreList}(t)$ 为 $\text{scoreCandidateRelationWords}(t)$ 的前 K 个元素
 - 3.3. 令集合 $\text{RelationWords}(t)$ 为 $\text{scoreList}(t)$ 和 IGList 的交集
 4. 返回集合 $\{\text{RelationWords}(t)|t \in \text{types}\}$
-

3.2.4 后处理

候选关系三元组集合中包含大量噪声，本节中使用关系指示词词表和句式规则来过滤这些噪声。同时还包含一些关系指示词抽取不完整的三元组，我们使用补全关系指示词的方法来解决这个问题。下面将分别介绍这些过滤和补全方法。

1) 使用关系指示词词表过滤三元组

候选三元组中的关系指示词包含很多噪声，例如从句子“陈曦主任近 6 年为佳木斯地区完成的部分首创手术”中抽出的候选关系三元组（陈曦，主任，佳木斯地区），这是由于候选关系三元组中的候选关系指示词包含很多不能指示关系的词语。在 3.2.3 小节中针对每一个实体对类型都生成了一个关系指示词词表，通过关系指示词词表可以过滤掉这些噪声。基于关系指示词词表过滤候选三元组的算法

如下表:

算法 3-2: 通过指示词词表过滤三元组

输入: 候选关系三元组集合 CandidateTriples , $\{\text{RelationWord}(t) | t \in \text{types}\}$

输出: 关系三元组集合 RelationTriples

步骤:

1. 初始化 RelationTriples 为空
 2. 对集合 CandidateTriples 中的每一个关系三元组 $\text{triple}(\text{entity1}, \text{relationWords}, \text{entity2})$
 - 2.1. 令 t 为: entity1 的实体类型- entity2 的实体类型
 - 2.2. 对 relationWords 中的每个词语 relationWord
 - 2.2.1. 如果 relationWord 不属于 $\text{RelationWord}(t)$, 那么:

把 relationWord 从集合 relationWords 中删除
 - 2.3. 如果集合 relationWords 非空, 那么:

把 triple 加入到集合 RelationTriples
 3. 返回 RelationTriples
-

2) 使用句式规则过滤三元组

从某些固定的句式抽取出来的三元组 $(i, \text{relationWords}, j)$ 很可能是噪声, 其中 i 是第一个实体在句子中的位置, j 是第二个实体在句子中的位置。下面是两条噪声句式。

➤ 系指示词包含动词且第二个实体后面第一个词语是动词, 其形式化描述为:

$$\text{hasV}(\text{relationWords}) \wedge \text{isV}(\text{pos}_{j+1}) \Rightarrow \text{isErrorTriple}(i, \text{relationWords}, j)$$

这类句式往往存在连动结构, 三元组无法描述其完整的关系实例。例如从“傅红雪告诉叶开说……”抽取的三元组(傅红雪, 告诉, 叶开)是不完整的, 本章的方法还无法处理这类句式, 所以将其过滤。

➤ 关系指示词都是名词且句子中第二个实体后面第一个词语是“的”, 其形式化描述为:

$$\neg \text{hasV}(\text{relationWords}) \wedge \text{isDE}(\text{word}_{j+1}) \Rightarrow \text{isErrorTriple}(i, \text{relationWords}, j)$$

这类句式正确抽取结果中存在关系的两个元素是第一个实体和第二个实体的所有格。例如从“宏仁集团的总裁是王泉仁的爸爸”抽取的三元组(宏仁集团, 总裁, 王泉仁)是一个错误的三元组, 其正确抽取结果是(宏仁集团, 总裁, 王泉仁的爸爸)。但本章只处理实体之间的关系, 所以过滤从这类句式抽取的三元组。

本文制定了句式过滤规则：如果三元组所在句子满足上述两种句式，那么三元组将被从候选集合中删除。

3) 补全关系指示词

在句子“<PER>王树国</PER>担任<ORG>哈尔滨工业大学</ORG>校长。”中，由于“校长”不是“PER-ORG”关系指示词词表中的词语，所以在补全关系指示词之前的关系三元组抽取结果是（王树国，担任，哈尔滨工业大学），显然这是一个错误的关系三元组，我们对这些错误进行处理，将“校长”补全到三元组的关系指示词中。

补全关系指示词主要针对实体对类型为 PER-LOC 和 PER-ORG 的关系三元组。对于实体对类型是 PER-LOC 的关系三元组，考察实体 2 右侧 3 个词语，如果发现某个词语在 LOC-PER 关系指示词词表中，那么把这个词语添加到关系三元组的关系指示词中。同理，对于实体对类型是 PER-ORG 的关系三元组，考察实体 2 右侧 3 个词语，如果发现某个词语在 ORG-PER 关系指示词词表中，那么把这个词语添加到关系三元组的关系指示词中。

3.3 实验结果及其分析

3.3.1 数据及评价方法

本章实验使用的网络文本语料抽取正文后共 10G 文本，网页包含以下三个来源：

- 1) 百度百科³160W 个网页
- 2) 新浪音乐新闻⁴（2008 年~2012 年）
- 3) 搜狗新闻语料⁵（2006 年，2012 年 6 月~2012 年 7 月）

为了评估句式过滤规则和补全关系指示词的效果，我们设置了两组不同的实验：

- 1) UnCORE：完整的系统。
- 2) UnCORE-post：UnCORE 除去句式规则过滤和补全关系指示词两个步骤后的系统。

对于网络文本上的关系三元组抽取结果很难直接评价召回率，所以使用三元组的数量来反映召回率。准确率的评价方法是对每种方法获取的每个实体对类型取结

³ <http://baike.baidu.com/>

⁴ <http://ent.sina.com.cn/music/roll.html>

⁵ <http://www.sogou.com/labs/dl/ca.html>

果中随机抽取 200 个关系三元组（共 2000 个关系三元组），然后人工判断每个关系三元组正确与否。

同时我们还在 Ontonotes4.0 上构建的开放式实体关系抽取语料上进行关系三元组抽取实验，从而评价方法的 P、R、F 值。由于 UnCORE 在获取关系指示词词表的时候是基于大规模候选三元组集合的，而 Ontonotes4.0 上的语料规模过小，无法获取大量的候选三元组。所以在 Ontonotes4.0 上实验时，本文使用从网络文本中挖掘的关系指示词词表。

3.3.2 结果及分析

本文对不同实验参数进行测试，发现参数设置如表 3-1 时，实验效果最好。

表 3-1 最优参数设置

Table 3-1 Value of Parameters

参数	N	K	maxDistance	maxEntityDistance	leftWordNumber	rightWordNumber
值	6000	5000	5	0	0	0

表 3-2 是从网络文本中抽取的各个实体对类型关系指示词词表中排名前 20 的词语，可以看出这些词语大多数都能描述实体之间的语义关系。当然也有一些噪声，如 LOC-PER 的关系指示词词表中的“雄鹰”，我们对关系指示词词表的抽取结果进行分析，发现这些错误大都是由于网络文本不规范和命名实体识别结果不准确而导致的。

表 3-2 从网络文本中抽取的各个实体对类型关系指示词词表的前 20 个关系指示词

Table 3-2 Top 20 relation words in each domain

实体对类型	关系指示词词表前 20 个关系指示词
LOC-PER	总统 选手 首相 市长 名将 作家 国务卿 省长 雄鹰 舞台 笔画 大使 诗人 科学家 物理学家 村民 数学家 国防部长 哲学家 国王
PER-LOC	出生 祖籍 离开 原籍 下台 率领 躬耕 生于 故里 南巡 病逝 访问 回到 追悼会 流放 统一 全家 遗体 走遍 来到
ORG-PER	主任 书记 局长 所长 秘书长 董事长 院长 部长 会长 主席 司长 委员长 总经理 总裁 研究员 执行官 科室 理事长 校长 总工程师
PER-ORG	现任 担任 做客 调任 哀思 代表 考入 致辞 出任 考上 毕业 当选 母校 杀人案 考取 辞去 加入 兼任 受聘 主持
PER-PER	妻子 儿子 女儿 饰演 弟弟 丈夫 扮演 哥哥 妹妹 遗孀 女友 母亲 夫人 父亲 扮演者 神似 好友 男友 女婿 长子

表 3-3 是在网络文本语料上抽取的关系三元组样例，句子中的斜体代表存在关系的两个实体，黑体代表关系指示词。关系三元组的评价结果如表 3-4 所示。图 3-4

是三元组抽取结果中正确关系三元组的数量，是一个估计值，其大小为三元组数量乘以准确率。

表 3-3 网络文本中抽取的关系三元组样例

Table 3-3 Samples of relation triples extrction

实体对类型	关系三元组	句子
LOC-PER	香港 导演 严浩	能说双语的 香港著名导演严浩也积极加盟。
	美国 总统 奥巴马	涨工资后，他的年薪是美国总统奥巴马的 5 倍。
PER-LOC	佟铁鑫 出生 辽宁锦州	男中音歌唱家佟铁鑫出生于辽宁锦州的一个音乐世家。
	秦始皇 统一 中国	秦始皇统一中国后，置齐地东部为琅琊郡，郡驻地在今天的琅琊镇。
ORG-PER	英特尔 公关经理 牛大鹏	英特尔公关经理牛大鹏并没有正面确认该信息。
	腾讯 董事长 马化腾	昨天，腾讯董事长马化腾在其微博上直接表态，重申腾讯不会做手机。
PER-ORG	林茨 效力 布拉加队	林茨目前效力于布拉加队，本赛季中前期表现出色。
	李开复 担任院长 微软亚洲研究院	上世纪 90 年代末，李开复曾担任微软亚洲研究院首任院长。
PER-PER	李冰冰 妹妹 李雪	李冰冰为妹妹李雪补办婚礼。
	奥多姆 经纪人 杰夫·施瓦茨	小牛已经给了奥多姆的经纪人杰夫·施瓦茨充分的自由去为奥多姆寻求下家。

表 3-4 网络文本上的关系三元组抽取结果

Table 3-4 Performance of relation triples extraction on the web data

实体对类型	三元组数量		准确率(%)	
	UnCORE-post	UnCORE	UnCORE-post	UnCORE
LOC-PER	289309	266080	72.00	78.00
PER-LOC	178734	110244	37.50	56.00
ORG-PER	211007	203318	95.00	99.00
PER-ORG	31574	18665	39.50	79.00
PER-PER	76498	35982	61.50	78.50
微平均			68.01	80.97

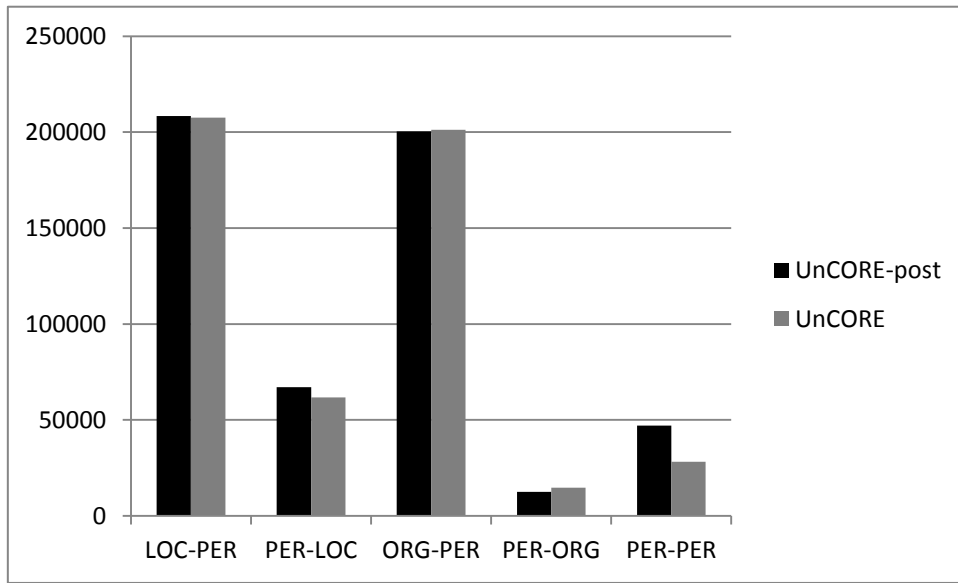


图 3-4 正确的三元组数目

Fig. 3-4 The number of corret relation triples

表 3-5 是在 Ontonotes4.0 的实验结果。由于 Ontonotes4.0 上的文本数量太少，无法使用本文的方法获取关系指示词词表，所以使用了在网络文本语料上构建的关系指示词词表。从表 3-5 中可以看出，UnCORE 的召回率比 UnCORE 的召回率有小幅下降，但是准确率的提升幅度是很明显的。

表 3-5 Ontonotes4.0 上的实验结果

Table 3-5 Performance of relation triples extraction on Ontonotes4.0

抽取方法	准确率(%)	召回率(%)	F 值(%)
UnCORE-post	69.19	50.20	58.18
UnCORE	77.18	48.55	59.61

通过实验结果分析，可以得出以下结论：

- 1) UnCORE 的微观平均准确率比 UnCORE-post 提高 12.96%，这说明句式过滤规则覆盖了大部分错误的关系三元组。
- 2) 使用句式规则和补全关系指示词后，PER-LOC 和 PER-PER 两个实体对类型的正确关系三元组数量下降较多，但是这两个实体对类型的关系三元组抽取准确率提高幅度很大，分别提高了 18.5% 和 17%。
- 3) PER-ORG 实体对类型的关系三元组抽取结果不但提高准确率，还增加了正确关系三元组的数量，其原因是在后处理中补全了关系指示词。通过补全关系指示词，可以从类似“PER 出任 ORG [职位]”的句式抽取正确的三元组 (PER, 出任[职位], ORG)。

- 4) 目前典型的开放式信息抽取系统 ReVerb 识别名词短语之间关系，其抽取结果最好的前 30%三元组准确率为 80%^[35]，UnCORE 的抽取结果的准确率在不排序的情况下达到 80%以上。
- 5) 在把方法移植到别的领域（Ontonotes4.0，新闻领域）时，三元组抽取的效果并没有太大的变化，这也证实了 UnCORE 的鲁棒性。

我们在不同的语料规模上进行实验，以评价语料规模对实验效果的影响。把包含候选三元组的句子集合 10 等份，然后设置 10 组对比实验。第一组实验使用一份候选集合，第二组实验使用 2 份候选集合，依次类推，知道第十组实验使用全部的候选集合。

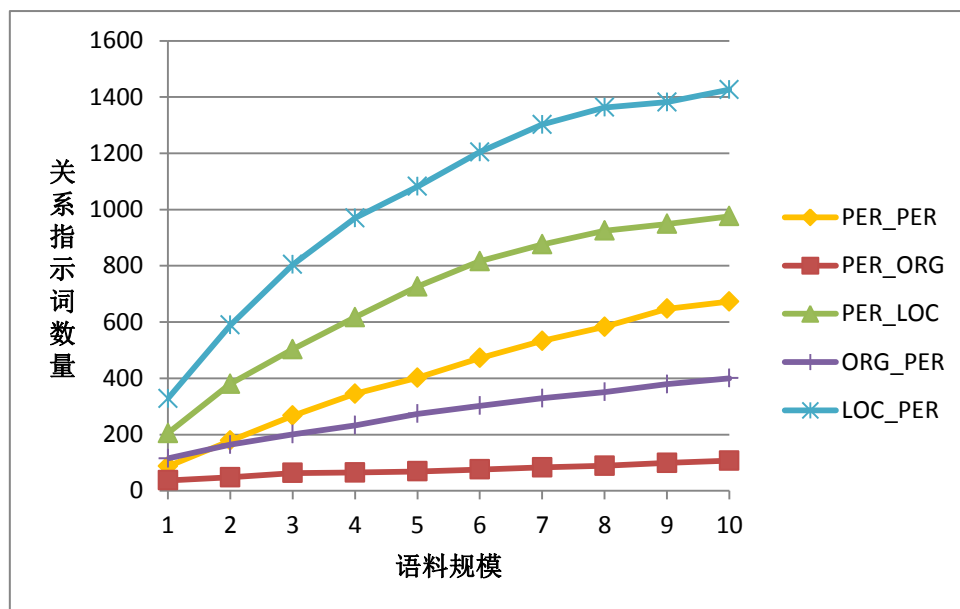


图 3-5 关系指示词数目随语料规模的变化趋势

Fig. 3-5 The relation between the number of relation words and corpus

图 3-5 显示，关系指示词的数目随着语料规模增大而增加，但是增长趋势有所减缓。这也说明互联网上的关系描述形式很丰富，很难通过人工构建一个全面的关系类型体系。

如图 3-6 所示，关系三元组数量随着语料规模的增大而增加，关系三元组数量的增大趋势一直很稳定。并且当增加一份语料时，关系三元组的增加数量要比从语料规模是 1 时抽取出来的三元组数量多，例如 PER-PER 实体对类型，从 2 份语料中抽取出来的关系三元组数量要比从 1 份语料中抽取出来的数量多 3651 个，多出来的数量这比单独从 1 份语料中出去出来的关系三元组数量(2640 个)大很多。这是由于在语料增大时，关系指示词的数量也越来越多，所以从单位语料中挖掘出的关系三元组数量将增多。

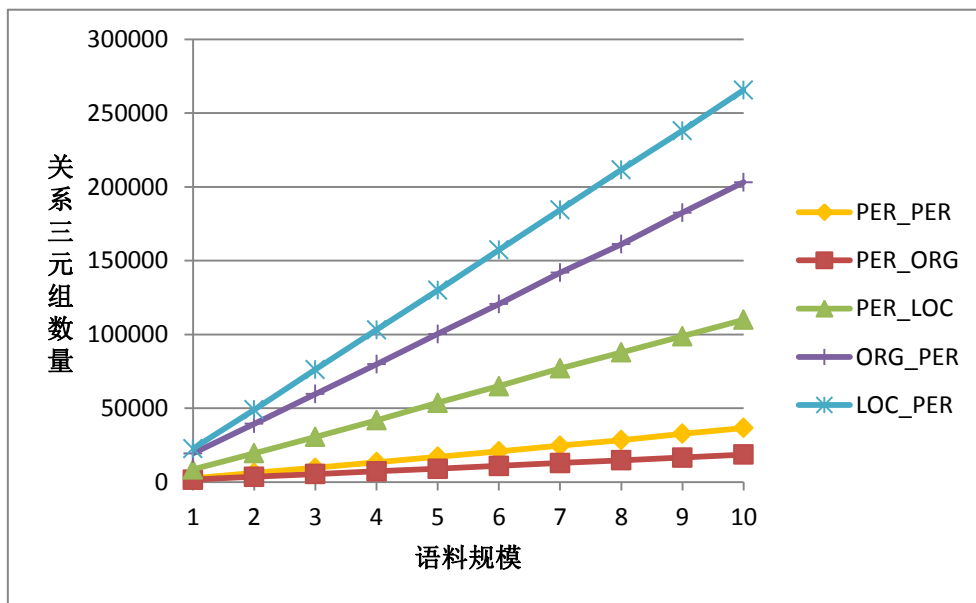


图 3-6 关系三元组数量随着语料规模的变化趋势

Fig. 3-6 The relation between the number of relation triples and corpus

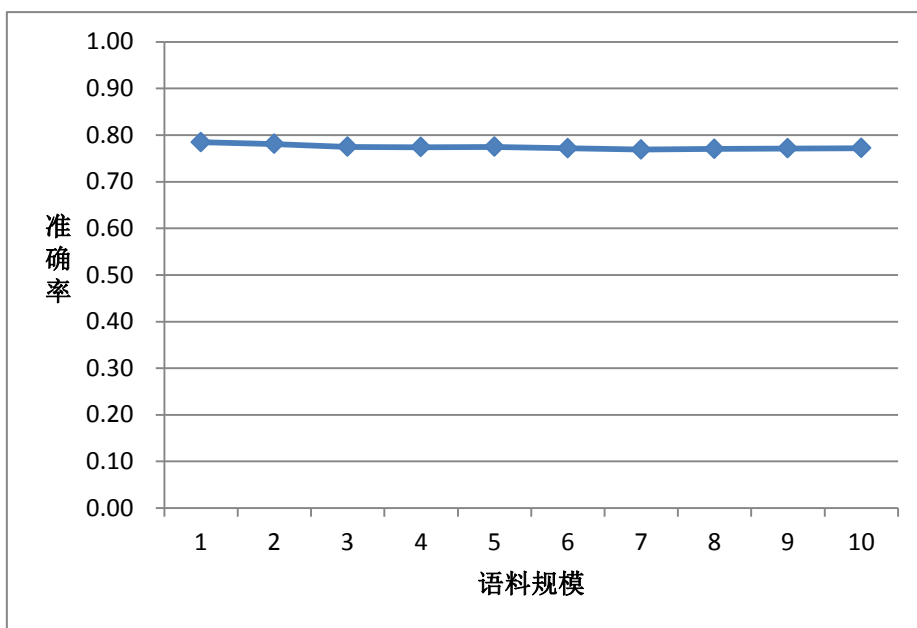


图 3-7 关系三元组准确率随着语料规模的变化趋势

Fig. 3-7 The relation between precision and corpus

在不同语料规模上，关系三元组抽取的准确率（Ontonotes4.0 上的测试结果）并没有太大的变化，如图 3-7 所示。说明了本章所提出的关系三元组抽取方法可以是很稳定的。

3.3.3 错误分析

我们通过分析错误的关系三元组发现，实体识别错误占很大的比例，如表 3-6 所示。实体错误会导致关系三元组抽取错误，例如句子“S O H O <LOC>中国</LOC>首席执行官<PER>张欣</PER>等中国民营企业家在会场发言或参与主题讨论。”中“S O H O 中国”是一个机构，但是命名实体识别出地名“中国”，从而导致抽取出来错误的三元组（中国，首席执行官，张欣）。

表 3-6 实体识别错误的三元组所占比例

Table 3-6 The percentage of triples with wrong entity

实体对类型	实体错误占有所有三元组的比例 (%)	实体错误所占错误三元组的组比例 (%)
LOC-PER	14.50	65.91
PER-LOC	20.00	45.45
ORG-PER	1.00	100.00
PER-ORG	4.00	19.05
PER-PER	12.50	58.14

关系指示词词表中包含一些不能指示关系的词语，这些词语被称为错误的关系指示词。这些错误的关系指示词可以分为两类：第一类是这些错误的关系指示词总出现在关系三元组的上下文中，一个典型的例子是“先生”总出现在 PER-ORG 的上下文中，从而被识别成 PER-ORG 的关系指示词，例如“<PER>黄如论</PER>先生荣获<ORG>国家民政部</ORG>颁发的爱心捐助奖个人奖项”；第二类错误是实体识别错误带来的影响，句子“<PER>菊科艾属</PER>植物<PER>冷蒿 Artemisia</PER>”中两个被识别成 PER 的实体都是错误的，并且这种情况非常多，所以导致“植物”被识别成一个关系指示词，进而影响最终的关系三元组抽取结果。

下面是对抽取结果中错误的三元组进行更细致的分类：

1) 实体边界错误

实体边界错误占很大一部分，例如<ORG>广州市第四中学<ORG>中只识别出<LOC>广州市<LOC>。

<LOC> 广 州 市 </LOC> 第 四 <relation_word> 中 学 校 长 助 理 </relation_word><PER>黄小燕</PER>黄小燕同志，广州市第四中学校长助理、党委委员，中学高级教师，广州市十佳青年语文教师，广州市高二语文中心组成员。

姜煜摄中新网上海6月17日电（记者姜煜）正在上海为“日本电影周”担任“亲善大使”的2012年国际小姐世界大赛<LOC>日本</LOC><relation_word>代表</relation_word><PER>吉松育</PER>美，17日对中新网记者表示，很想在中国有所发展。

2) 分词错误

分词错误对关系指示词识别影响较大，下面例子中“中后卫”应该是一个词语，但是被分成两个词。

本报讯（记者刘超峰）由于中后卫外援一直是河南建业足球队急需的“人才”，中原球迷也在热切盼望实力派中后卫的加盟，昨天，建业引进中后卫外援的消息在网上传开效力于瑞士锡永队的<LOC>巴西</LOC>中<relation_word>后卫</relation_word><PER>阿代尔顿</PER>，已经与建业签约，近两天将宣布加盟。

3) 关系指示词抽取不全

关系指示词抽取不全主要是由于正确的结果中包含修饰成分，例如“副”、“常务”、“前”等。

市县两级200亿元财政资金重点投向十大产业……昨天，杭州市政协举行“以创新促转型做强实体经济”专题常委会，<LOC>杭州市</LOC><relation_word>常务副</relation_word>市长</relation_word><PER>杨戌标</PER>通报了杭州市实体经济发展情况。

4) 实体类型错误

故事背景封印着一千八百年前三国时代无法完成国家统一理想的英雄魂魄的<PER>勾玉</PER><relation_word>辗转</relation_word><relation_word>流落</relation_word>到现代的<LOC>日本国</LOC>，并散落关东各地。

5) 指示词识别错误

针对库班地区将发生第二轮洪灾的传言，俄罗斯紧急情况部南部地区中心新闻处负责人奥列格·格列科夫日前表示，<LOC>库班</LOC><relation_word>地区</relation_word>的<LOC>克雷姆斯克</LOC>不可能发生第二轮洪灾，所有水库均运行正常，他还驳斥了这种挑衅性言论。

3.4 本章小结

本章提出面向互联网的无指导开放式中文实体关系抽取方法，首先使用实体之间的距离限制和关系指示词的位置限制获取候选关系三元组，然后采用全局排序

和类型排序的方法来挖掘关系指示词，最后使用关系指示词和句式规则对关系三元组进行过滤。在获取大量关系三元组的同时，还保证了 80% 以上的微观平均准确率。把方法应用于 Ontonotes4.0 时，关系三元组抽取的准确率变化并不太大，这证实了 UnCORE 方法有较强可移植性。同时，我们还在不同语料规模上做实验，发现关系三元组的数量随着语料规模的增大呈线性增长，并且关系三元组抽取的准确率一直很稳定。

UnCORE 不需要预先标注语料库，只需要输入大规模的文本，在挖掘大规模关系三元组的同时，还可以得到的文本中包含的关系指示词。当语料规模较大时，从中挖掘出来的关系指示词数目和类型将非常丰富。相对于有指导的方法，UnCORE 不包含时间复杂度高的算法，所以可以把本方法用于对时间复杂度要求高的应用中。本章为从大规模文本中挖掘出关系实例提供了一套快速有效的方法。

第4章 开放式中文实体关系类型体系自动构建

4.1 引言

通过观察和分析关系指示词集合，我们发现有部分关系指示词表达相同或相近的实体关系，例如“PER-PER”实体对类型的关系指示词词表中包含“妻子”、“老婆”、“丈夫”等描述“夫妻关系”的词语。为了把这些描述关系相同或相近的关系指示词聚集在一起，我们将提出基于关系指示词聚类的方法来自动构建开放式实体关系类型体系。

我们以特定的实体对类型的关系指示词词表（第三章中的处理结果）为处理对象，然后使用不同的相似度计算方法（基于 HowNet^[39]、基于 RNN-LM^[40]），最后通过不同的聚类算法（层次聚类算法，近邻传播算法^[41]）对关系指示词进行聚类。最终形成一个关系类型体系，聚类结果中的每一个簇就是一类实体关系。

4.2 基于聚类的开放式实体关系类型体系自动构建

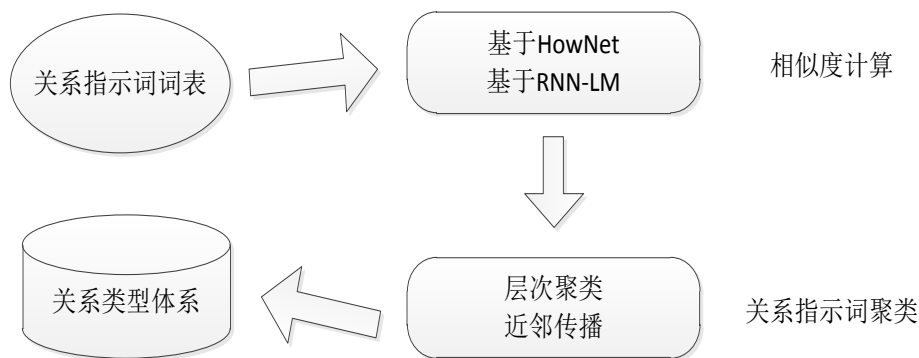


图 4-1 基于聚类的开放式实体关系类型体系自动构建

Fig. 4-1 architecture of relation types building

基于关系指示词聚类的开放式实体关系类型体系自动构建算法框图如图 4-1 所示。从图中可以看出，算法的输入是关系指示词词表，输出是关系类型体系。算法包括两个主要的步骤：相似度计算和关系指示词聚类。下面我们将详细介绍这两个步骤。

4.2.1 相似度计算

为了获取更好的性能，我们尝试两种不同的相似度计算方法：基于 HowNet 的相似度计算和基于 RNN-LM 的相似度计算。下面我们将详细介绍这两种相似度计算方法。

1) 基于 HowNet 的相似度计算

HowNet 又称为《知网》^[39]，有两个主要的概念：“概念”和“义原”。义原是对概念的最小描述单位，一个词语可以表示若干个概念。HowNet 与普通的语义词典不同，HowNet 试图使用一系列“义原”来描述“概念”。通过 HowNet 中的“义原”，我们可以计算不同词之间的语义相似度。

$$\text{sim}(rw_i, rw_j) = \frac{2 * C_{ij}}{C_i + C_j} \quad (4-1)$$

公式(4-1)描述了关系指示词之间的语义相似度计算方法，其中 rw_i 和 rw_j 是要计算语义相似度的两个词语， C_i 表示 rw_i 的概念定义中义原的个数， C_j 表示 rw_j 概念定义中义原的个数， C_{ij} 表示 rw_i 和 rw_j 的概念定义中相同义原的数目。借助 HowNet，通过公式(4-1)可以计算任意两个关系指示词的语义相似度，并且两个关系指示词相同的义原个数越多，相似度值越大。在对关系指示词进行聚类时，我们使用语义相似度作为关系指示词之间的相似度。

2) 基于 RNN-LM 的相似度计算

对于在 HowNet 中的两个关系指示词，我们可以通过公式（4-1）计算他们的相似度，但是如果有一个关系指示词不在 HowNet 中，我们将无法精确计算他们之间的相似度。为了克服这个问题，我们使用 RNN-LM^[40]（Recurrent Neural Network Language Model）训练模型，使用一个实数向量来描述关系指示词。然后通过余弦相似度来计算两个关系指示词之间的相似度。

4.2.2 聚类

我们尝试了两类不同的聚类方法来对关系指示词进行聚类，分别是：层次聚类和近邻传播算法。下面我们将分别介绍这两种算法。

1) 层次聚类算法

由于贪心策略不同，层次聚类的实现可以分为两种：自底向上（bottom-up）和自顶向下（top-down）。在实现聚类算法时，我们采用自底向上的策略，首先初始

化每一个关系指示词为一个单独的簇，然后每次把最相似的两个簇合并，直到簇的数目少于阈值为止。

传统的层次聚类算法的输出时一个层次化的聚类结果，但是我们对关系指示词进行聚类的时候，并不需要输出层次化的聚类结果，所以针对我们的问题，对传统的层次聚类算法有所改进。

算法 4-1: 改进的层次聚类算法

输入: 关系指示词词表 $\text{RelationWords}=\{\text{rw}_1 \dots \text{rw}_k\}$

聚类数目 n

关系指示词集合之间相似度计算函数 $f(c_i, c_j)$

输出: n 个簇

步骤:

1. 初始化 $c_i = \{\text{rw}_i\}$
 2. $C = \{c_i \mid 1 \leq i \leq k\}$
 3. $j = k+1$
 4. 循环 $k - n$ 次
 - 4.1. $(c_{i1}, c_{i2}) = \arg \max_{(c_u, c_v) \in C \times C} f(c_u, c_v)$
 - 4.2. $c_j = c_{i1} \cup c_{i2}$
 - 4.3. $C = C \setminus \{c_{i1}, c_{i2}\} \cup c_j$
 5. 返回 C
-

算法 4-1 是改进后的层次聚类算法，和原始层次聚类算法的不同点在于步骤 4，原始算法中循环 $k-1$ 次，最后所有的元素都被聚集在一个集合中。改进后，元素将被分成 n 个集合。

在算法 4-1 的输入中的关系指示词集合相似度计算函数 $f(c_i, c_j)$ 不同于 4.2.1 节中描述的关系指示词之间的相似度， $f(c_i, c_j)$ 计算的是两个关系指示词集合之间的相似度，他需要满足单调性：

$$\forall c, c', c'' \subseteq C \text{ 满足 } \min(f(c, c'), f(c, c'')) \geq f(c, c' \cup c'')$$

关系指示词集合之间相似度计算函数单调性保证在聚类过程中不会增加关系指示词之间的相似程度。否则会出现这样的情况：原本相似度很小的两个关系指示词，在经过若干步合并操作后，两个关系指示词之间的相似度变大了。这显然是错误的。

针对层次聚类算法中的关系指示词集合之间相似度计算函数，我们设计了两种

方案：

- 单连通：关系指示词集合之间的相似度是两个集合间最相似的两个关系指示词的相似度。
- 全连通：关系指示词集合之间的相似度是两个集合间最不想死的两个关系指示词的相似度。

不难证明，通过单连通和全连通的集合设计的相似度计算函数 $f(c_i, c_j)$ 满足单调性。

2) 近邻传播算法

Frey B J^[41]于 2007 年提出近邻传播 (AP, Affinity Propagation, 简称 AP) 算法。AP 算法不需要预先给定类别数目，这不同于 k-means 聚类。AP 聚类算法是在元素的相似度矩阵基础上，通过信息传播进行聚类的。它把每一个聚类元素都看作一个潜在的聚类中心，并且不要求元素之间的相似度矩阵对称（即允许 $\text{sim}(a, b) \neq \text{sim}(b, a)$ ）。代表矩阵 R 和适选矩阵 A 是近邻传播算法的两个重要参数。矩阵 R 中的元素 $r(i, k)$ (responsibility) 衡量使用聚类元素 x_k 作为聚类元素 x_i 所属类别中心的能力，矩阵 A 中的元素 $a(i, k)$ (availability) 衡量当 x_k 作为类别中心的时候 x_i 属于这个类的程度。 $r(i, k)$ 与 $a(i, k)$ 的和越大，说明聚类元素 k 越适合作为聚类中心点出现，且聚类元素 i 越可能属于这个类别。AP 算法经过若干次迭代后，会把相似的关系指示词聚成一个集合，并且每个集合都有一个类别中心（能代表该集合的一个关系指示词），满足元素 i 隶属的簇的中心点为 $\arg \max_k (r(i, k) + a(i, k))$ 。

4.3 实验结果及其分析

4.3.1 数据与评价标准

实验数据为第三章生成的 PER-PER 实体对类型的关系指示词集合。为了评估自动构建的关系类型体系是否合理，我们人工构建了一个标准的评价集，由于聚类标准的评价方法不一，我们选取了两种不同的评价方法。下面我们将详细介绍标准评价集的构建方法和聚类记过的评价方法。

1) 标准评价集

由于 PER-PER 实体对类型的关系指示词包含 600 多个关系指示词，很难对所有的关系指示词进行细致的分类，并且随着语料规模的增长，关系指示词的数目

还会不断的变化，所以我们在 PER-PER 的关系指示词集合中随机（频次越高的关系指示词被选中的概率越大）抽取了 97 个关系指示词进行人工分类。表 4-1 是我们构建的标准评价集。

表 4-1 中包含两列：第一列是人工给定的关系类型，这是关系指示词集合能描述的关系；第二列是能描述特定关系类型的关系指示词。我们在构建评价集时，尽量使各个实体关系类型内部的关系指示词内聚性高，同时，对于表述相关系类型的关系指示词，必须被分到同一个关系类型，例如“男朋友”和“男友”描述的是同一类关系。

表 4-1 关系类型体系评价集
Table 4-1 Evaluation set of relation types

关系类型	关系指示词
兄弟姐妹	兄长 妹妹 表哥 胞弟 弟弟 姐姐 大哥 哥哥 兄弟 姐妹 双胞胎
亲属	儿媳 祖父 祖母 侄女 岳父 舅舅 孙女 奶奶 孙子 姑姑 婆婆 外甥 侄子 爷爷 侄儿 叔叔 女婿
子女	小儿子 父子 大儿子 儿子 养女 长子 长女 女儿 次子
情侣	男朋友 男友 恋情 女友 初恋 恋人 情人 约会
夫妻	未婚夫 老婆 老伴 夫人 新婚 新娘 太太 爱人 丈夫 未婚妻 妻子
前任夫妻	离婚 前妻
老乡	同乡 老乡
同门	前辈
师徒	徒弟 教练 老师 恩师 师傅 弟子 班主任
接班人	传人 接班人 继承人
同事	同事
好友	老友 老朋友 友人 好友 朋友 战友 挚友 队友
偶像	粉丝 偶像 模仿
助手	助手
扮演者	饰演 扮演 主演 扮演者
经纪人	经纪人
父母	父亲 妈妈 母亲 爸爸
酷似	酷似
暗恋对象	暗恋
合作搭档	制作人 搭档

2) 评价标准

由于我们的标准评价集并不全部包含 PER-PER 实体对类型的关系指示词，而实验是针对真个关系指示词词表进行聚类的，所以我们需要一种策略使用标准集来评价聚类结果。我们的处理方法如下：

- 对聚类结果中的每一个簇，如果簇中的关系指示词不在标准评价集中，那么就把这个关系指示词从簇中删除。
- 如果簇中不包含关系指示词，那么把簇删除。

通过上述处理后，我们再使用两种不同的聚类评价方法进行评价：纯度和 F 值测度。这两种评价方法可以从不同的侧面对聚类结果进行评价。下面将分别描述着两种评价指标。

纯度（Purity）是衡量聚类结果的内聚性，考察的是聚类之后在同一个簇中的元素是否来自于标准评价集中同一个簇，即聚类结果的混乱程度，纯度越高聚类结果的混乱程度越低。纯度的计算方法如公式 4-3 所示。

$$\text{purity}(r) = \frac{1}{c_r} \max\{|c_r \cap c_i|\} \quad (4-2)$$

$$\text{avgPurity} = \sum_{r=1}^k \frac{|c_r|}{n} \text{purity}(r) \quad (4-3)$$

c_r 聚类结果中的簇， c_i 是标准集中的簇， k 是聚类结果中簇的个数， n 是关系指示词的数目。

从公式 4-3 可以看出，当聚类结果中每一个关系指示词是一个簇的时候， avgPurity 的值将为 100%，这说明纯度指标无法惩罚聚类算法中聚类倒退（即当个关系指示词是一个簇）的情形。F 值测度可以解决这个问题，它是准确率和召回率的加权调和平均数，如公式 4-7 所示，在这里，我们设置准确率和召回率的权值是一样的。

$$R(i, r) = \frac{|c_r \cap c_i|}{|c_i|} \quad (4-4)$$

$$P(i, r) = \frac{|c_r \cap c_i|}{|c_r|} \quad (4-5)$$

$$F(i, r) = \frac{2 \times R(i) \times P(i)}{R(i) + P(i)} \quad (4-6)$$

$$\text{avgF} = \sum_i \frac{|c_i|}{n} \max_r F(i, r) \quad (4-7)$$

4.3.2 结果与分析

我们设置了随机聚类算法作为 baseline，其做法是：给定的聚类数目 n ，对每一个关系指示词随机从 1 到 n 之间选取一个数作为类别标签。然后针对不同的聚类数目 n ，选取 F 值最高的一次作为 baseline。表 4-2 列出了 7 种关系指示词聚类的方法。

表 4-2 实验方法
Table 4-2 Methods of building relation types

方法	相似度计算方法	聚类算法
Baseline	无	随机
HowNet+AP	HowNet	AP
HowNet+HAC(single link)	HowNet	HAC(single link)
HowNet+HAC(complete link)	HowNet	HAC(complete link)
RNN-LM+AP	RNN-LM	AP
RNN-LM+ HAC(single link)	RNN-LM	HAC(single link)
RNN-LM+ HAC(complete link)	RNN-LM	HAC(complete link)

AP 聚类算法需要预先设置阈值参数 threshold，为了选取最好的实验结果，我们对 threshold 的不同进行了实验，参数 threshold 对实验效果的影响如图 4-2 所示。

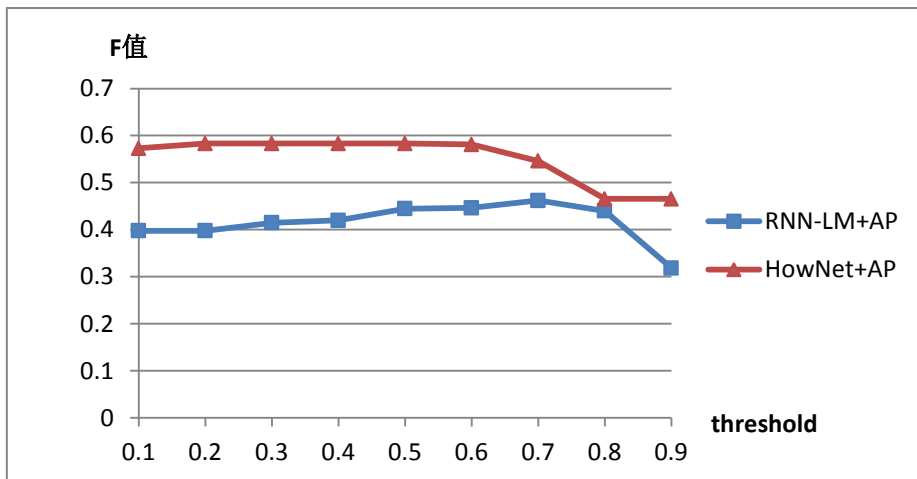


图 4-2 AP 算法中参数 threshold 对效果的影响

Fig. 4-2 The relation between performance and threshold

从图 4-2 中可以看出，基于 HowNet 计算相似度的聚类效果要比基于 RNN-LM 计算相似度的聚类效果好，这是由于 HowNet 包含了关系指示词的语义信息。RNN-LM+AP 方法的 F 值最高达到 44.65%，HowNet+AP 方法的 F 值最高达到 58.3%。

在层次聚类中，需要输入聚类数目，所以，我们尝试了不同聚类数目 m ，并且统计了 m 对聚类效果的影响，如图 4-3 所示。不难看出，基于全连通的层次聚类算法要比单连通的层次聚类算法效果好，这是由于全连通在计算两个关系指示词集合的相似度时，考虑了全局信息，而单连通只考虑了局部信息。

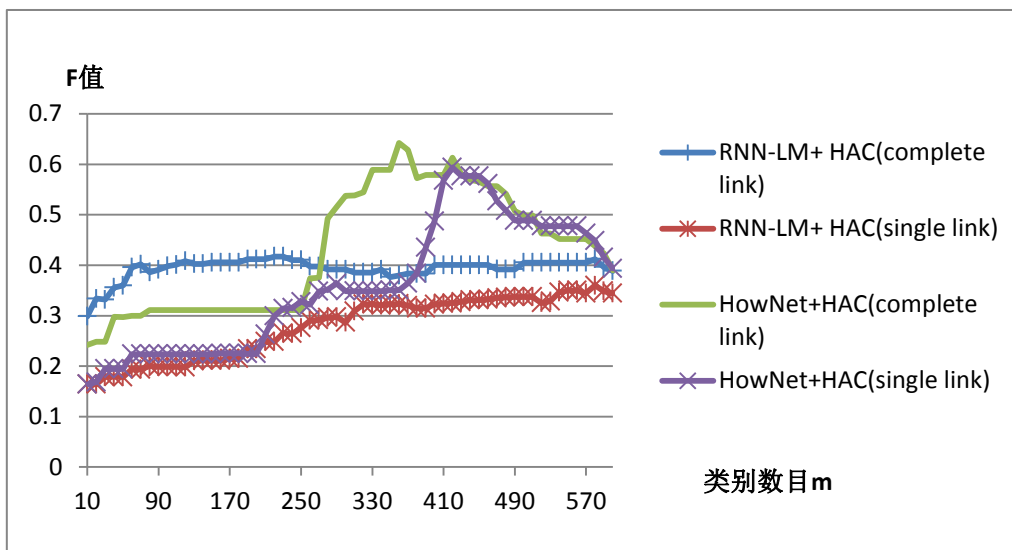


图 4-3 层次聚类参数对聚类效果的影响

Fig. 4-3 The relation between performance and cluster number

表 4-3 呈现了各个方法最好的效果 (F 值), 可以看出使用 HowNet 计算相似度要比使用 RMM-LM 计算相似度的效果好, 这是由于 HowNet 是一部语义资源, 相比之下 RMM-LM 是基于无指导的方法计算出来的。同时可以看出基于 AP 聚类算法的效果要比 HAC 算法好。在所有的方方法中, HowNet+HAC(complete link)取得了最优的效果, 其 F 值达到 64.25%, 这是由于 HowNet 引入了语义信息, 并且能覆盖全部的关系指示词。

表 4-3 各个方法的效果比较

Table 4-3 The performance of each method

方法	类别数目	纯度 (%)	F 值 (%)
baseline	540	91.75	33.43
RNN-LM+AP	266	65.98	44.65
RNN-LM+ HAC(complete link)	230	62.89	41.72
RNN-LM+ HAC(single link)	580	83.51	35.97
HowNet+AP	325	84.54	58.30
HowNet+HAC(complete link)	360	78.35	64.25
HowNet+HAC(single link)	420	85.57	59.37

4.4 本章小结

本章通过对关系指示词进行聚类而自动构建关系类型体系, 首先以 PER-PER 实体对类型的关系指示词集合为处理对象, 然后分别使用 HowNet 和 RNN-LM 的

方法来计算关系指示词之间的相似度，最后通过层次聚类算法或近邻传播算法对关系指示词进行聚类，其聚类结果就是 PER-PER 领域的关系类型体系。我们对各种方法进行对比，发现 HowNet+ HAC(complete link)的方法达到了最好的实验结果，其 F 值达到 64.25%。

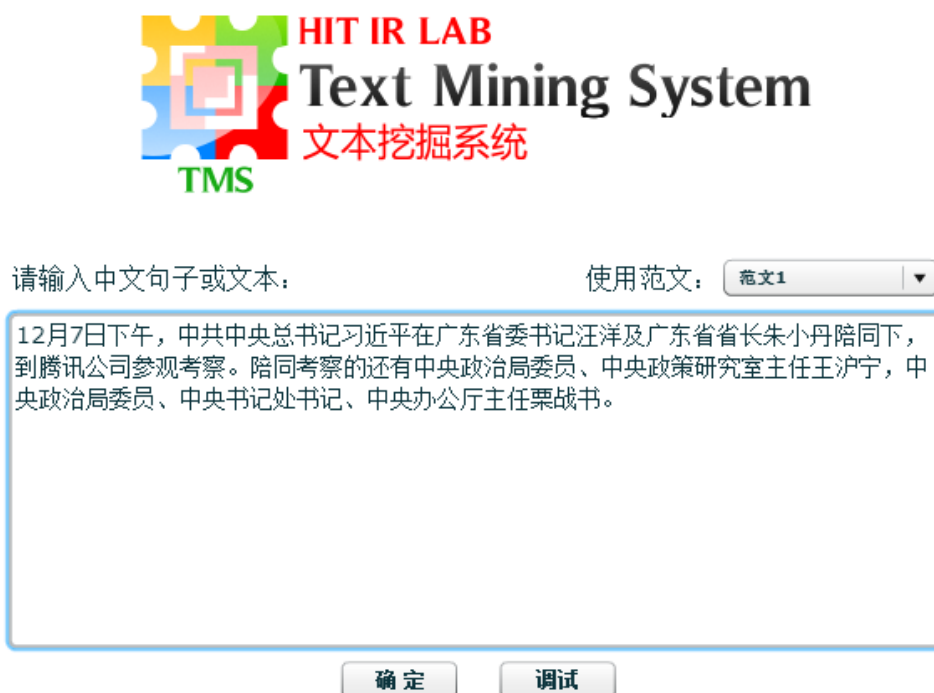
综上所述，针对开放式实体关系抽取任务，我们可以通过聚类的方法自动构建关系类型体系，可以为关系三元组的相似度计算、关系指示词归一化的相关研究提供参考。

第5章 开放式中文实体关系抽取平台设计与实现

5.1 引言

第二章中，本文提出了基于有指导的方法从句子中抽取关系三元组，在此基础上搭建了面向句子的开放式实体关系抽取系统。为了查询从互联网的大量网络文本中快速挖掘关系三元组，本文在第三章中提出了基于无指导的开放式实体关系抽取方法，并且使用该方法获取了大量的关系三元组，在此基础上，搭建了开放式实体关系三元组查询系统。

5.2 句子级开放式实体关系抽取系统



HIT IR LAB
Text Mining System
文本挖掘系统
TMS

请输入中文句子或文本： 使用范文： 范文1 ▼

12月7日下午，中共中央总书记习近平在广东省委书记汪洋及广东省省长朱小丹陪同下，到腾讯公司参观考察。陪同考察的还有中央政治局委员、中央政策研究室主任王沪宁，中央政治局委员、中央书记处书记、中央办公厅主任栗战书。

确定 调试

图 5-1 输入界面

Fig. 5-1 The interface of input

本节以第二章的方法为基础，搭建了一个句子级的开放式中文实体关系抽取系统⁶。前台使用 FLEX 对抽取结果进行展示，后台使用 JAVA 编程语言实现实体关

⁶ <http://ir.hit.edu.cn/opentms/>

系的抽取过程，通过 Tomcat 服务器实现前台和后台的通信。图 5-1 是系统的输入界面。

前台把输入文本传给后台，在通过断句、分词和命名实体的基础上，系统会对每一个句子进行开放式实体关系三元组抽取，并把抽取结果返回给前台。对于输入文本“12月7日下午，中共中央总书记习近平在广东省省委书记汪洋及广东省省长朱小丹陪同下，到腾讯公司参观考察。陪同考察的还有中央政策研究室主任王沪宁，中央办公厅主任栗战书。”的开放式实体关系抽取结果展示界面如图 5-2 所示。

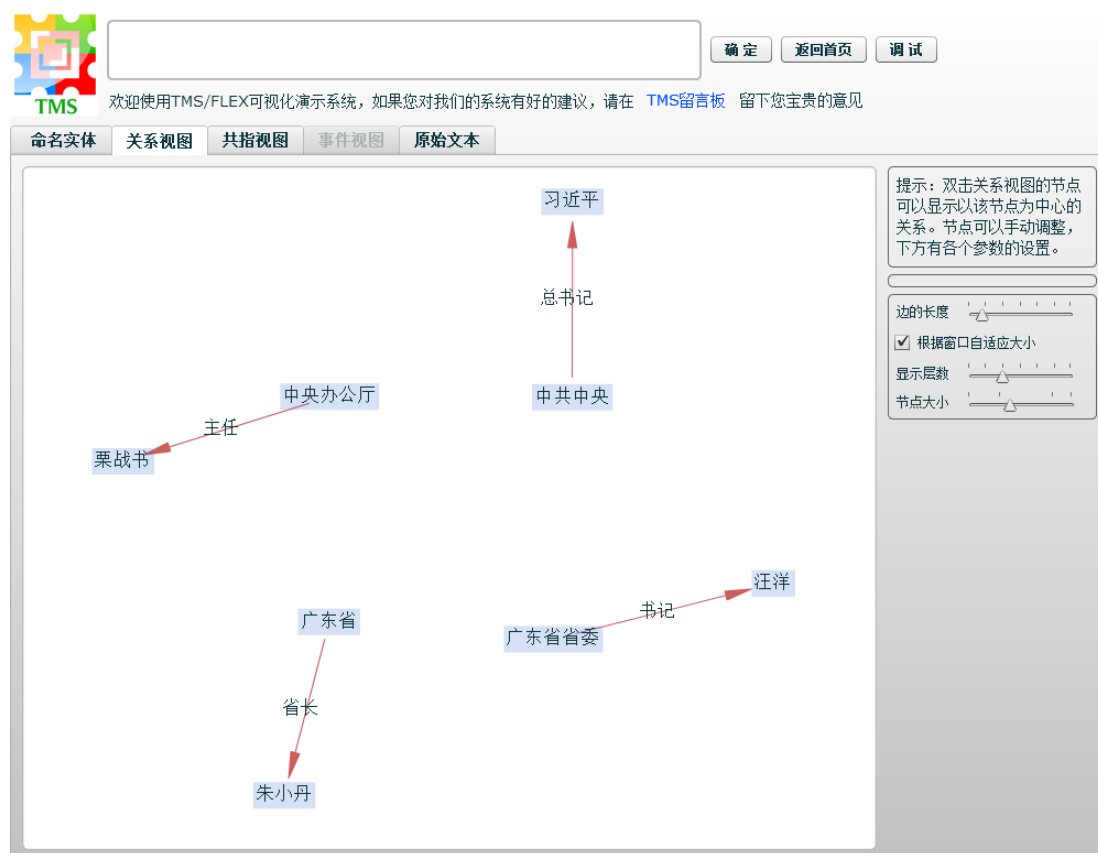


图 5-2 句子级的开放式实体关系抽取结果展示

Fig. 5-2 The result screen of relation extraction

从图 5-2 中可以看出，系统从文本中抽取出 4 个关系三元组（中共中央，总书记，习近平）、（广东省省委，书记，汪洋）、（广东省，省长，朱小丹）、（中央办公厅，主任，栗战书）。在演示界面中，存在关系的两个实体使用一个单向箭头相连，连线上会附有关系指示词。

5.3 开放式实体关系三元组查询系统

基于第三章的方法，我们从互联网文本中获取了大量的关系三元组。在此，我们提供一个关系三元组检索的演示系统⁷，用户输入一个实体，系统返回包含该实体的关系三元组，返回结果中还包含关系三元组所在的句子。系统的输入结果如图 5-3 所示。



图 5-3 查询实体输入界面

Fig. 5-3 The interface of input

图 5-4 是实体“哈尔滨工业大学”的查询结果，返回结果中每一行是一个关系三元组实例，前三列是关系三元组的信息，最后一列是关系三元组出现的句子。第四列是关系三元组的“可信度”，通过计算关系三元组出现的不同句子数目得到。

实体1	实体2	指示词	可信度	句子
哈尔滨工业大学	王树国	校长	2	省及哈尔滨市领导徐泽洲、林锋、夏杰、孙亮、于莎燕、张建星、宋希斌， 哈尔滨工业大学校长王树国 等。
哈尔滨工业大学	王树国	校长	2	，山东大学校长展涛在哈尔滨工业大学与 哈尔滨工业大学校长王树国 共同签署《山东大学——哈尔滨工业大学校企合作框架协议》。
马德骥	哈尔滨工业大学	毕业	1	资料显示， 马德骥 1982年 毕业于哈尔滨工业大学 ，曾就职于长春汽车研究所、一汽-大众、北亚集团等单位。
进岭	哈尔滨工业大学	毕业	1	李进岭 先生 毕业于哈尔滨工业大学 ，主修市场营销专业。
杨谦	哈尔滨工业大学	现任教授	1	杨谦 博士 现任哈尔滨工业大学教授 、博士生导师，黑龙江省植物病理学会副理事长，教育部生物技术与生物工程专业教学指导委员会委员，中国生物工程杂志专业评委、理事，哈尔滨工业大学大学生物工程专业学科带头人。
张管生	哈尔滨工业大学	毕业	1	张管生 教授 毕业于哈尔滨工业大学 ，是国内资深知名能源专家。
张曙	哈尔滨工业大学	毕业	1	张曙 教授 毕业于哈尔滨工业大学 。
哈尔滨工业大学	田太藤	教授	1	近日， 哈尔滨工业大学教授田太藤 关于文理科的观点引发争议。
哈尔滨工业大学	马军任	教授	1	清华大学环境科学与工程系教授王占生、 哈尔滨工业大学教授马军任 主持人。
哈尔滨工业大学	韩杰才	校长	1	签约仪式前，李群在青岛宾馆会见了 哈尔滨工业大学副校长韩杰才 一行。
哈尔滨工业大学	郝雨	学生	1	这一方面的音乐作品由 哈尔滨工业大学的学生郝雨 的说唱flash作品《大学自习室》。
哈尔滨工业大学	海王基金奖	博士生	1	获得1997年度 哈尔滨工业大学博士生海王基金奖 。
哈尔滨工业大学	沈世钊	教授	1	经中国工程院院士、 哈尔滨工业大学教授沈世钊 等10位业内知名专家评审认定，该项技术成果达到国际先进水平。
哈尔滨工业大学	民商法教研部	法学院	1	哈尔滨工业大学法学院民商法教研部主任
哈尔滨工业大学	梁迎春	研究所	1	在机械加工方面， 哈尔滨工业大学精密工程研究所梁迎春 教授将微细铣削技术应用到微型零件加工中，孙涛教授等人使用原子力显微镜实现微齿轮的加工，并成功刻画出微齿轮。
哈尔滨工业大学	杨苏	教授	1	1、 哈尔滨工业大学教授杨苏 学术任职
哈尔滨工业大学	张桓	工程学院	1	毕业于 哈尔滨工业大学汽车工程学院 的 张桓 ，有着比较扎实的计算机基础，进而把他的公司引入网络营销领域，也是顺其自然的事情。
哈尔滨工业大学	刘家琦	教授	1	哈尔滨工业大学教授刘家琦 1.北京医学院附属医院医生刘家琦
哈尔滨工业大学	刘哲	校长	1	霍虹桥由中铁铁路局理事、 哈尔滨工业大学校长刘哲 题字命名。
哈尔滨工业大学	冯纯伯	教研室	1	任 哈尔滨工业大学自动控制教研室冯纯伯 主任，1962年晋升为副教授。
冯仲云	哈尔滨工业大学	出任校长	1	冯仲云 出任 哈尔滨工业大学 校长，在他的努力下，许多烈士遗孤得到了最妥善的安置和最好的教育。

图 5-4 实体“哈尔滨工业大学”关系三元组查询结果

Fig. 5-4 The result screen of retrieval

同一个关系三元组出现的句子数目多，其被认可的程度也越高，所以，在展示关系三元组查询结时，我们使用可信度对其进行降序排序，使得可信度高的关系三元组排在前面，以提高用户体验。

⁷ <http://ir.hit.edu.cn/iknow/index.jsp>

5.4 本章小结

在前面章节的相关研究成果的基础上，本章设计并实现了“面向句子的开放式实体关系抽取系统”和“开放实体关系三元组查询系统”。在面向句子的开放式实体关系抽取系统中，用户输入文本，系统把从文本中抽取的开放式实体关系三元组清晰的展现出来。在开放式实体关系三元组查询系统中，用户输入实体，系统返回该实体相关的关系三元组，并且使用计算关系三元组出现的句子数目给出了关系三元组的可信度。

结 论

实体关系是描述实体之间语义关系的重要途径。实体关系抽取是信息抽取任务中的重要环节，也有着广泛的应用前景。随着 Web2.0 的迅猛发展，人们对实体关系抽取提出了新的要求，以适应从快速增长的海量互联网文本中迅速准确地获取对用户有价值的信息。传统的实体关系抽取需要预先定义关系类型体系，然而定义一个全面的实体关系类型体系是很困难的。开放式实体关系抽取技术通过使用关系指示词描述关系的方法解决了预先定义关系类型体系的问题，但是在中文上的研究还比较少。本文从开放式实体关系语料建设开始，对中文的开放式实体关系抽取进行了系统的研究。

本文的主要创新点和贡献包括以下几个方面：

(1) 本文针对中文的开放式实体关系抽取任务制定了语料规范，并且构建了 1000 篇文档的语料库。我们认真地分析语料中的语言现象，把开放式实体关系抽取任务分成两个子问题：实体对识别和关系指示词识别。针对两个子问题的解决先后顺序，分别设计了两种不同的解决方案：先识别实体对的方案和先识别关系指示词的方案。为了增强模型的移植能力，我们设计了泛化能力较强的特征：使用词性、实体的词序列等特征。对两种不同的方案进行实验，关系组抽取结果的 F 值达到 61.41%。

(2) 为了快速处理互联网上的海量文本，本文提出面向互联网的无指导开放式中文实体关系抽取方法，首先使用实体之间的距离限制和关系指示词的位置限制获取候选关系三元组，然后采用全局排序和类型排序的方法来挖掘关系指示词，最后使用关系指示词和句式规则对关系三元组进行过滤。该方法在获取大量关系三元组的同时，还保证了 80% 以上的微观平均准确率。我们还在不同的领域使用 UnCORE 方法，取得了较好的效果，证实了 UnCORE 方法的鲁棒性。

(3) 通过观察和分析关系指示词集合，我们发现有部分关系指示词表达相同或相近的实体关系，例如“PER-PER”实体对类型的关系指示词词表中包含“妻子”、“老婆”、“丈夫”等描述“夫妻关系”的词语。为了把这些描述关系相同或相近的关系指示词聚集在一起，我们将提出基于关系指示词聚类的方法来自动构建开放式实体关系类型体系。我们以特定的实体对类型的关系指示词集合为处理对象，然后使用不同的相似度计算方法计算关系指示词之间的相似度，通过 HAC 算法和 AP 算法对关系指示词进行聚类，最终形成一个类型丰富的关系体系。

尽管已经取得了一定的阶段性成果，但是还存在许多需要改进的地方以及值得研究的问题，有如下几部分：

（1）面向句子的开放式实体关系抽取方法的召回率还有提高的空间，将来可以采用句法特征来优化方法的召回率。

（2）优化命名实体识别效果。命名实体识别效果对关系三元组抽取任务有很大的影响，传统的命名实体识别方法在应用到网络文本上时，其效果会有较大的下降。将来可以针对关系抽取任务优化命名实体识别的效果。

（3）关系指示词推理。我们对实体之间的语义关系使用关系指示词来描述，同样关系指示词之间也存在语义关系。例如关系三元组（A，父亲，B）和（B，父亲，C）可以推理出（A，爷爷，C）。将来，在大规模的关系三元组的基础上，自动学习这种推理关系。

参考文献

- [1] 车万翔, 刘挺, 李生. 实体关系自动抽取. 中文信息学报. 2005, 19(2):1-6.
- [2] ACE. Annotation guidelines for entity detection and tracking. ACE2004. 2004.
- [3] C. Plake, T. Schiemann, M. Pankalla, J. Hakenberg, and U. Leser. Alibaba: PubMed as a Graph. Bioinformatics, vol. 22. 2006:2444-2445.
- [4] T. S. Jayram, R. Krishnamurthy, S. Raghavan, S. Vaithyanathan, and H. Zhu. Avatar Information Extraction System. IEEE Data Engineering Bulletin, vol. 29. 2006:40-48.
- [5] F. M. Suchanek, G. Ifrim, and G. Weikum. Combining Linguistic and Statistical Analysis to Extract Relations from Web Documents. In KDD06: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2006:712-717.
- [6] B. M. Suchanek, G. Kasneci and G. Weikum. Yago: A core of semantic knowledge. In WWW07: Proceedings of the 16th International Conference on World Wide Web. 2007:698-706.
- [7] E. Agichtein and L. Gravano. Snowball: Extracting Relations from Large Plain-text Collection. In proceedings of the 5th ACM International Conference on Digital Libraries. 2000:85-94.
- [8] Michele Banko, Michael J Cafarella, Stephen Soderland, Matt Broadhead and Oren Etzioni. Open Information Extraction from the Web. In IJCAI. 2007:2670-2676.
- [9] Chinchor, N. Overview of MUC-7/MET-2. In Message Understanding Conference Proceedings: MUC-7.
- [10] Ralph Grishman and Beth Sundheim. Message Understanding Conference-6: A Brief History. In proceedings of the 16th Conference on Computational Linguistics. 1996:466-471.
- [11] In Proceedings of the Seventh Message Understanding Conference (MUC-7)[C]. National Institute of Standards and Technology. 1998.
- [12] Chinchor N., Marsh E. MUC-7 Information Extraction Task Definition. In proceeding of Seventh Message Understanding Conference. 1998:2-3.
- [13] ACE. Automatic Content Extraction 2008 Evaluation Plan (ACE08). In

- proceedings of the ACE 2008 Evaluation. 2008:1-16.
- [14] J. Aitken. Learning Information Extraction Rules: An Inductive Logic Programming Approach. In proceedings of the 15th European Conference on Artificial Intelligence. 2002:355-359.
- [15] D. McDonald, H. Chen, H. Su and B. Marshall. Extracting Gene Pathway Relations using A Hybrid Grammar: The Arizona Relation Parser. *Bioinformatics*, vol. 20. 2004:3370-3378
- [16] W. Shen, A. Doan, J. F. Naughton and R. Ramakrishna. Declarative Information Extraction using Datalog with Embedded Extraction Predicates. In *VLDB*. 2007:1033-1044.
- [17] J. Jiang and C. Zhai. A Systematic Exploration of the Feature Space for relation Extraction. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*. 2007:113-120.
- [18] 董静, 孙乐、冯元勇. 中文实体关系抽取中的特征选择研究. *中文信息学报*. 2007, 21(4):80-91.
- [19] Yee Seng Chan and Dan Roth. Exploiting background knowledge for relation extraction. In *Proceedings of the 23th International Conference on Computational Linguistics (COLING 10)*. Association for Computational Linguistics 2012:152-160.
- [20] I. Tsoukantaridis, T. Joachims, T. Hofmann and Y. Altun. Large Margin Methods for Structured and Interdependent Output Variables. *Journal of Machine Learning Research (JMLR)*. 2005:1453-1484.
- [21] R. C. Bunescu and R. J. Mooney. A Shortest Path Dependency Kernel for Relation Extraction. In *HLT05: Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*. 2005:724-731.
- [22] A. Culotta and J. Sorensen. Dependency Tree Kernels for Relation Extraction. In *Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL04)*. 2003:24-31.
- [23] M. Wang. A Re-examination of Dependency Path Kernels for Relation Extraction. In *Proceedings of INCNLP 2008*. 2008:841-846.
- [24] D. Zelenko, C. Aone and A. Richardella. Kernel Methods for Relation Extraction. *Journal of Machine Learning Research*, vol.3. 2003:1083-1106.

- [25] M. Zhang, J. Zhang, J. Su and G.D. Zhou. A Composite Kernel to Extract Relations between Entities with both Flat and Structured Features. In Proceedings of the 21th International Conference on Computational Linguistics and 44th Annual Meeting of the Association of Computational Linguistics(COLING/ACL-2006), Sydney, Australia. 2006:825-832.
- [26] S. Zhao and R. Grishman. Extracting Relations with Integrated Information using Kernel Methods. In ACL05: Proceedings of the 43th Annual Meeting on Association for Computational Linguistics. 2005:419-426.
- [27] E. Agichtei. Extracting Relations from Large Text Collections. PhD thesis, Columbia University. 2005.
- [28] R. Bunescu and R. Mooney. Learning to Extract relations form the Web using Minimal Supervision. In Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics. 2007:576-583.
- [29] B. Rosenfeld and R. Feldman. Using Corpus Statistics on Entities to improve Semi-supervised Relation Extraction form the Web. In Proceedings of the Association of Computational Linguistics. 2007:600-607.
- [30] Y. Shinyama and S. Sekine. Preemptive Information Extraction using Unrestricted Relation Discovery. In HLT-NAACL. 2006:304-311.
- [31] P. D. Turney. Expressing Implicit Semantic Relations without Supervision. In ALC. 2006:313-320.
- [32] Yulan Yan, Naoaki Okazaki, Yutaka Natsuo, Zhenglu Yang and Mitsuru Ishizuka. Unsupervised Relation Extraction by Mining Wikipedia Texts Using Information from the Web. In Proceedings of 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP. 2009:1021-1029.
- [33] Fei Wu, Daniel S. Weld. Open information extraction using Wikipedia. ACL '10 Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. 2010:118-127.
- [34] Surdeanu M, Tibshirani J, Nallapati R, et al. Multi-instance multi-label learning for relation extraction. Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. 2012: 455-465.
- [35] Anthony Fader, Stephen Soderland, Oren Etzioni. EMNLP '11 Proceedings of the Conference on Empirical Methods in Natural Language Processing. 2011:

1535-1545.

- [36] Yao L, Riedel S, McCallum A. Unsupervised relation discovery with sense disambiguation. Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics. 2012: 712-720.
- [37] 王莉峰. 领域自适应的中文实体关系抽取研究. 哈尔滨: 哈尔滨工业大学硕士论文, 2011.
- [38] Che W, Li Z, Liu T. LTP: A Chinese Language Technology Platform. In Proceedings of the Coling 2010: Demonstrations. 2010.08:13-16.
- [39] Dong Z, Dong Q. HowNet and the Computation of Meaning. World Scientific Publishing Co., Inc., 2006.
- [40] Mikolov T, Karafiát M, Burget L, et al. Recurrent neural network based language model. Proceedings of Interspeech. 2010: 1045-1048.
- [41] Frey B J, Dueck D. Clustering by passing messages between data points. science, 2007, 315(5814): 972-976.

攻读硕士学位期间发表的论文及其它成果

（一）发表的学术论文

- [1] 刘安安, 秦兵, 刘挺. 无指导的开放式中文实体关系抽取[C]. 第十九届全国信息检索学术会议, 2013.

哈尔滨工业大学学位论文原创性声明和使用权限

学位论文原创性声明

本人郑重声明：此处所提交的学位论文《开放式中文实体关系抽取研究》，是本人在导师指导下，在哈尔滨工业大学攻读学位期间独立进行研究工作所取得的成果，且学位论文中除已标注引用文献的部分外不包含他人完成或已发表的研究成果。对本学位论文的研究工作做出重要贡献的个人和集体，均已在文中以明确方式注明。

作者签名：日期：年 月 日

学位论文使用权限

学位论文是研究生在哈尔滨工业大学攻读学位期间完成的成果，知识产权归属哈尔滨工业大学。学位论文的使用权限如下：

(1) 学校可以采用影印、缩印或其他复制手段保存研究生上交的学位论文，并向国家图书馆报送学位论文；(2) 学校可以将学位论文部分或全部内容编入有关数据库进行检索和提供相应阅览服务；(3) 研究生毕业后发表与此学位论文研究成果相关的学术论文和其他成果时，应征得导师同意，且第一署名单位为哈尔滨工业大学。

保密论文在保密期内遵守有关保密规定，解密后适用于此使用权限规定。

本人知悉学位论文的使用权限，并将遵守有关规定。

作者签名：日期：年 月 日

导师签名：日期：年 月 日

致 谢

农历癸巳年仲夏之月，值此论文即将完成之际，心中感慨良多。论文能以顺利完成，我既体会到辛勤劳动后的喜悦，又深感它与大家的帮助和支持是分不开的。

感谢哈工大社会计算与信息检索研究中心的所有老师，特别感谢研究中心主任刘挺教授，感谢您为我们提供优越的工作环境和良好的学习科研氛围。您开阔的视野、敏锐的思维、严谨的学风，以及严以律己、宽以待人的高尚品质无不是我学习的楷模。

感谢导师秦兵教授一直以来对我的信任和鼓励。秦老师在生活上给了我无微不至的关心，在研究上给了我自由发挥的空间，让我在很多项目的研究和开发中锻炼动手和管理能力，使我学习了知识，开阔了视野，相信这些将使我终生受益。

感谢已经毕业的王莉峰和胡燊师兄，是你们把我带到了 NLP、IR、IE 领域，与你们相处的日子里我学到了很多，成长了很多，你们广阔的视野，出色的研发能力和团队合作意识给我留下了深刻的印象，并一直影响、改变着我。

感谢付瑞吉、李正华、宋巍师兄，和你们一起度过的每个“羽毛球之夜”是那么的激情。

感谢 TM 组所有组员，感谢一起学习奋斗过的 11 级 SCIRer（宋原[TM]、张健[UA]、赵江江[TM]、邓知龙[LA]、陆子龙[SN]、焦扬[SN]、王沛[TM]、慕福楠[UA]、刘飞[UA]）以及实验室其他成员，谢谢你们平日里热心的帮助、信任和鼓励，希望你们学习、工作顺利。

感谢我的两位好朋友、好兄弟吴峰和邓本洋，是你们在我空虚寂寞是陪伴我吃喝玩乐。

感谢哈工大对我的培养，希望母校蒸蒸日上，培养出更多优秀的人才，为国强民富作出更瞩目的贡献。

感谢我那还未出现的女友，谢谢你的矜持使得我能顺利完成论文。

感谢养育我长大成人的家人，谢谢你们始终如一的关心和支持，这些都是我不断向前进取的重要动力和保障。

感谢所有曾经给予我关心、支持和帮助的人们，愿你们好运常伴！