无指导的开放式中文实体关系抽取*

刘安安，秦兵，刘挺

哈尔滨工业大学计算机科学与技术学院，黑龙江哈尔滨，150001

E-mail: {aaliu, bqin, tliu}@ir.hit.edu.cn

摘　要：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。

关键词：开放式实体关系抽取；无指导；关系三元组；关系指示词

Unsupervised Chinese Open Entity Relation Extraction

Anan Liu, Bing Qin, Ting Liu

School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001

E-mail: {aaliu, bqin, tliu}@ir.hit.edu.cn

Abstract: 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. 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 candidate relation triples. Finally UnCORE filters candidate relation triples by using the extracted relation words and some sentence rules. Results show that UnCORE extracts large scale relation triples at precision higher than 80%.

Keywords: Open Entity Relation Extraction; unsupervised; relation triple; relation word

1　引言

实体关系抽取的目的是发现和识别实体之间的语义关系[1]，是信息抽取的重要环节。传统的实体关系抽取方法需要预先确定关系类型体系，然而预先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[2]使用实体上下文中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。其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 是上下文中描述实体之间语义关系的词或词序列。例如在文本“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。”中可以抽取出关系三元组（腾讯，首席执行官，马化腾）。

英文的开放式实体关系抽取相关研究已经比较成熟。Michele Banko 等人[2]提出 TextRunner 系统，利用启发式规则在宾州树库中自动构建语料，然后训练模型识别关系三

*本课题受到国家自然科学基金(61122012, 61273321)及博士点基金（20122302110039）资助.

元组。Fei Wu 等人[3]提出 WOE 系统，巧妙地使用维基百科中信息框（Infobox）的内容来标注语料，这种方法提高了训练语料的质量和数量。Mihai Surdeanu 等人[4]认为同一个实体对在不同的句子中呈现出不同的关系，从而提出了 MIML 模型提高自动标注语料的准确率。Anthony Fader 等人[5]对 TextRunner 系统和 WOE 系统的抽取结果进行分析，发现错误的关系三元组主要分为不合逻辑和无意义两类，从而提出了先识别关系指示词的 ReVerb 系统。Limin Yao 等人[6]提出了基于 LDA 的关系模板聚类方法构建关系类型体系。

中文的开放式实体关系抽取相关研究还比较少。中文和英文的语言现象相差较大，所以无法把英文上的方法直接移植到中文上。王莉峰[7]提出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，在音乐领域人与人之间的关系上取得不错的效果。

本文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，本文提出一种新颖的无指导开放式实体关系抽取方法，主要研究人、机构、地点之间的实体关系开放式描述。

2 面向大规模网络文本的无指导开放式中文实体关系抽取

如图 1 所示，UnCORE 共包含 4 个模块：预处理模块，生成候选三元组模块，生成关系指示词词表模块，后处理模块。

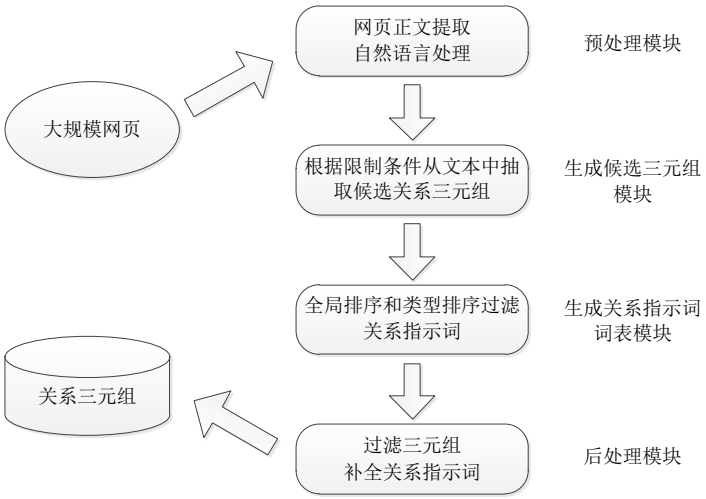

图 1 面向大规模网络文本的开放式实体关系抽取模型

Figure 1 Architecture of UnCORE

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：

(1) 网页正文提取：使用基于文本行分布的正文抽取[1]方法抽取网页中的正文文本。

(2) 自然语言处理：使用 LTP[8]对网页文本进行断句、分词、词性标注和命名实体识别。

[1] https://code.google.com/p/cx-extractor/

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，我们标注了一个开放式关系抽取语料，用来统计关系三元组分布规律。通过分析语料，本文提出了两个生成候选关系三元组的限制条件。

2.2.1. 实体之间的距离限制

图 2 上点（5，0.7457）表示两个实体之间词数目小于等于 5 的关系实例数目占总关系三元组数目的 74.57%。从图 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。也就是说词距离小的实体之间更可能存在关系，因此，在生成候选关系三元组的时候，规定两个实体之间词的数目不超过 maxDistance。

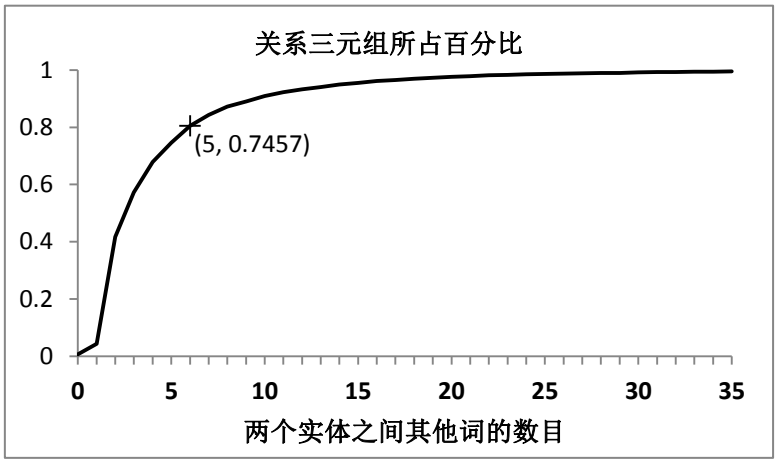

图 2 关系三元组数目在词距离上的分布情况

Figure 2 The distribution of relation triples with different word distance

从图 3 中可以得出和图 2 类似的结论，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，规定实体之间其他实体数量不能超过 maxEntityDistance。

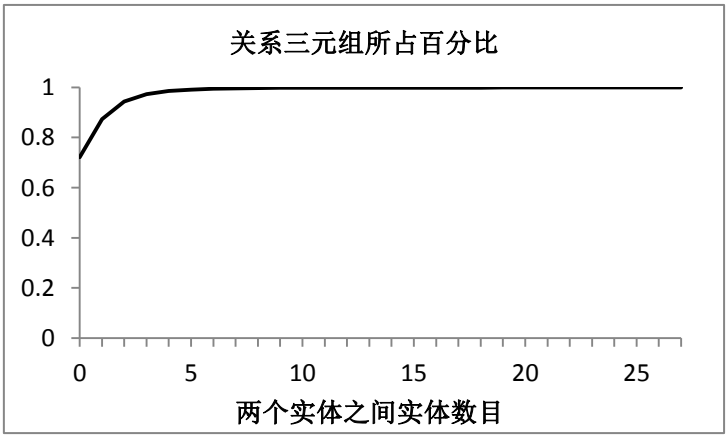

图 3 关系三元组数目在实体距离上的分布

Figure 3 The distribution of relation triples with different entity distance

2.2.2. 关系指示词的位置限制

如表 1 所示，包括 93.6%的关系实例都能在原始文本中找到一些词语来描述实体之间的语义关系，这证实了使用三元组来描述一个关系实例是可行的。

表 1 指示词在句子中的位置分布

Table 1 The distribution of relation triples with the position of relation word

关系指示词的位置	关系实例数目（个）	比例（%）
两个实体之间	3177	75.36
第二个实体右边	609	14.44
第一个实体左边	160	3.80
没有指示词	240	6.40

为了提高候选关系三元组的准确率，本文提出了关系指示词位置限制：在生成候选关系三元组时，把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 个名词和动词作为候选关系三元组的候选关系指示词。

2.3 生成关系指示词词表

候选关系指示词集合中包含了大量的噪声，所以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。

2.3.1. 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(1)所示。

$$IG(rel) = H(\text{types}) - H(types|\text{rel}) \quad (1)$$

其中 rel 是候选关系指示词，在实验中发现与人相关的实体对类型的关系指示词比较丰富，所以实验中关注的实体对类型为 types={PER-PER, PER-ORG, PER-LOC, ORG-PER,LOC-PER}，我们将区分两个实体的先后顺序。

2.3.2. 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(2)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。

$$\text{score}(\text{rel}, t) = p(t|\text{rel}) \log c(rel, t) \quad (2)$$

公式(2)要求指示词 rel 要在实体对类型 t 上出现概率高，且 rel 和 t 共现次数多，才能使$\text{score}(\text{rel}, t)$的值大。具体地，“总裁”在 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，并在 ORG-PER 中出现的频率很高，从而使 score(总裁|ORG-PER)的值大。

2.3.3. 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 1：生成关系指示词词表
输入：候选关系指示词集合 CandidateRelationWords，IG(rel)，score(rel, t)，types 输出：关系指示词词表{RelationWords(t)|t∈types} 步骤： 1. 令集合 IGCandidateRelationWords 为 CandidateRelationWords 按照 IG(rel)值降序排序结果 2. 令集合 IGList 为 IGCandidateRelationWords 的前 N 个元素 3. 对集合 types 中的每个元素 t 3.1. 令集合 scoreCandidateRelationWords(t)为 CandidateRelationWords 按照score(rel, t)值降序排序结果 3.2. 令集合 scoreList(t)为 scoreCandidateRelationWords(t)的前 K 个元素 3.3. 令集合 RelationWords(t)为 scoreList(t)和 IGList 的交集 4. 返回集合{RelationWords(t)|t∈types}

2.4 后处理

候选关系三元组集合中包含大量噪声，本文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，本文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。

2.4.1. 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（陈曦，主任，佳木斯地区）。通过关系指示词词表可以过滤掉这些噪声：针对每个实体对类型的候选三元组，过滤掉不在词表中的候选关系指示词；如果过滤后的三元组不包含关系指示词，将此三元组从候选集合中删除。

2.4.2. 句式规则过滤三元组

从某些固定的句式中抽取出来的三元组（i, 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。

(1) 关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：

$$hasV(\text{relationWords}) \wedge isV(pos_{j+1}) \Longrightarrow isErrorTriple(i, relationWords, j)$$

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。

(2) 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：

$$\neg hasV(relationWords) \wedge isDE(word_{j+1}) \Longrightarrow isErrorTriple(i, relationWords, j)$$

例如从“宏仁的总裁是王泉仁的爸爸”抽取出错误的三元组（宏仁，总裁，王泉仁）。

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。

2.4.3. 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，所以抽取出错误的三元组（王树国，

担任，哈尔滨工业大学）。本文对这类错误进行处理，将缺失的关系指示词补全到三元组中。

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。对实体对类型是 PER-ORG 的关系三元组做类似的处理。

3 实验

3.1 实验数据及评价方法

本文实验使用的网络文本语料抽取正文后共 10G 文本，包含以下三个来源：

(1) 百度百科[2]160W 个网页

(2) 新浪音乐新闻[3]（2008 年~2012 年）

(3) 搜狗新闻语料[4]（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-post：UnCORE 除去句式规则过滤和补全关系指示词两个步骤

(2) UnCORE：完整的系统

对于网络文本上的关系三元组抽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个实体对类型从其抽取结果中随机抽取 200 个关系三元组，然后人工判断每个关系三元组正确与否。

3.2 实验结果及分析

本文测试了不同参数的实验结果，发现参数设置为（N=6000，K=5000，maxDistance=5，maxEntityDistance=0，leftWordNumber=0，rightWordNumber=0）时，实验效果最好。

表 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关系。

表 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 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

[2] http://baike.baidu.com/

[3] http://ent.sina.com.cn/music/roll.html

[4] http://www.sogou.com/labs/dl/ca.html

表 3 是在网络文本语料上抽取的关系三元组样例，评价结果如表 4 所示。图 4 是识别正确的关系三元组数目，这是一个估计值，其大小为三元组数量乘以准确率。

表 3 网络文本中抽取的关系三元组样例

Table 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。
PER-LOC	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。
PER-ORG	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。

表 4 网络文本上的关系三元组抽取结果

Table 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	289,309	266,080	72.00	78.00
PER-LOC	178,734	110,244	37.50	56.00
ORG-PER	211,007	203,318	95.00	99.00
PER-ORG	31,574	18,665	39.50	79.00
PER-PER	76,498	35,982	61.50	78.50
微平均			68.01	80.97

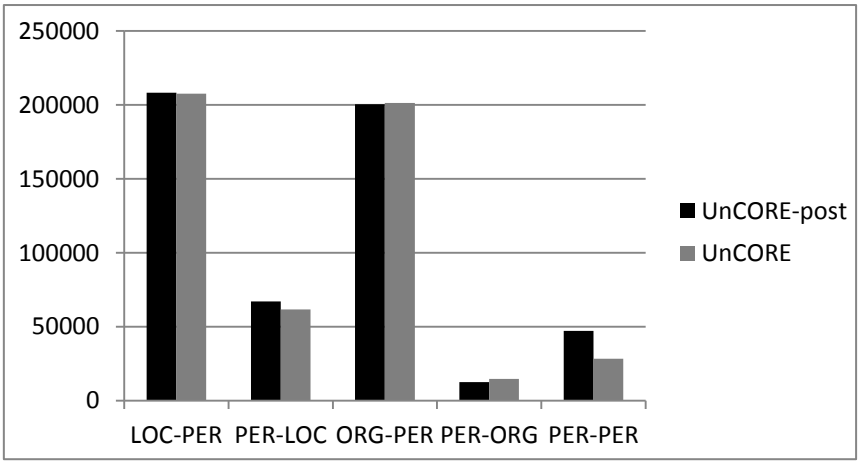

图 4 正确的三元组数目

Figure 4 The number of corret relation triples

通过实验结果分析，可以得出以下结论：

(1) UnCORE 的微平均准确率比 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 的正确关系三元组数量下降较多，但是这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了准确率，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]”的句式中抽取正确的三元组（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%[5]， UnCORE 的抽取结果准确率在达到 80%以上。

实体识别错误多实验结果影响较大，如表 5 所示。实体错误会导致关系三元组抽取错误，例如句子“ＳＯＨＯ<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“ＳＯＨＯ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 5 实体识别错误的三元组所占比例

Table 5 The percentage of triples with wrong entity

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

4 结论及未来工作

本文提出了一种面向大规模网络文本的无指导中文实体关系抽取方法，可以有效地从文本中抽取关系三元组，其微平均准确率达到 80%以上。

我们将尝试在更大规模的语料上做实验，探索实验效果与语料规模之间的关系；不同的关系指示词可能描述同一类关系，例如“爸爸”和“父亲”都可以描述“父子”关系，我们将探索如何自动构建一个丰富的实体关系类型体系；命名实体识别的结果对关系三元组的抽取效果影响很大，我们将为关系抽取任务优化命名实体识别效果。

参 考 文 献

[1] Chinchor N, Marsh E. MUC-7 Information Extraction Task Definition. In proceeding of Seventh Message Understanding Conference MUC-7, Appendices. 1998:359-367.

[2] Banko M, Cafarella M J, Soderland S, et al. Open Information Extraction from the Web. In IJCAI. 2007:2670-2676.

[3] Wu F, Weld D S. Open information extraction using Wikipedia. ACL '10 Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. 2010:118-127.

[4] 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.

[5] Fader A, Soderland S, Etzioni O. Identifying Relation for Open Information Extraction. Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2011: 1535-1545.

[6] 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.

[7] 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，2011.

[8] Che W, Li Z, Liu T. LTP: A Chinese Language Technology Platform. In Proceedings of the Coling 2010:Demonstrations. 2010.08:13-16.