计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.2015.20131550
52(5):1029-1035,2015

无指导的中文开放式实体关系抽取

秦 兵 刘安安 刘 挺
(哈尔滨工业大学计算机科学与技术学院 哈尔滨 150001)
(bqin@ir.hit.edu.cn)

Unsupervised Chinese Open Entity Relation Extraction

Qin Bing, Liu An'an, and Liu Ting
(*School of Computer Science and Technology, Harbin Institute of Technology, Harbin* 150001)

Abstract Entity relation extraction is an important task in information extraction which helps people find knowledge quickly and accurately in various text. Traditionally, entity relation extraction methods require a pre-defined set of relation types and a corpus with manual tags. But it is difficult to build a well-defined architecture of the relation types and it takes a lot of time to label a corpus. Open entity relation extraction is the task of extracting relation triples from natural language text without pre-defined relation types. There is a lot of research in the field of English open entity relation extraction, but rarely in the field of Chinese open entity relation extraction. This paper presents the UnCORE (unsupervised Chinese open entity relation extraction method for the Web). UnCORE is an unsupervised open entity relation extraction method which discovers relation triples from large-scale Web text. UnCORE exploits using word distance and entity distance constraints to generate candidate relation triples from the raw corpus, 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%.

Key words open entity relation extraction; unsupervised; relation triple; relation word; information extraction

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.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;信息抽取

中图法分类号 TP391

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(*entity*1, *relationWords*, *entity*2),其中(*entity*1,

收稿日期:2013-11-29;**修回日期**:2014-05-06
基金项目:国家自然科学基金项目(61122012,61273321);国家"八六三"高技术研究发展计划基金项目(2012AA011102)

*entity*2)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. 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元组. Wu 等人[3]提出 WOE(wikipedia-based open extractor)系统,巧妙地使用维基百科中信息框的内容来标注语料,这种方法提高了训练语料的质量和数量. 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(multi-instance multi-label)模型提高自动标注语料的准确率. 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,从而提出了先识别关系指示词的 ReVerb 系统. 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.

本文通过分析中文关系抽取语料库,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.

1 无指导的中文开放式实体关系抽取

如图 1 所示,面向大规模网络文本的无指导中文开放式实体关系抽取模型(unsupervised Chinese open entity relation extraction method for the Web, 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及后处理模块.

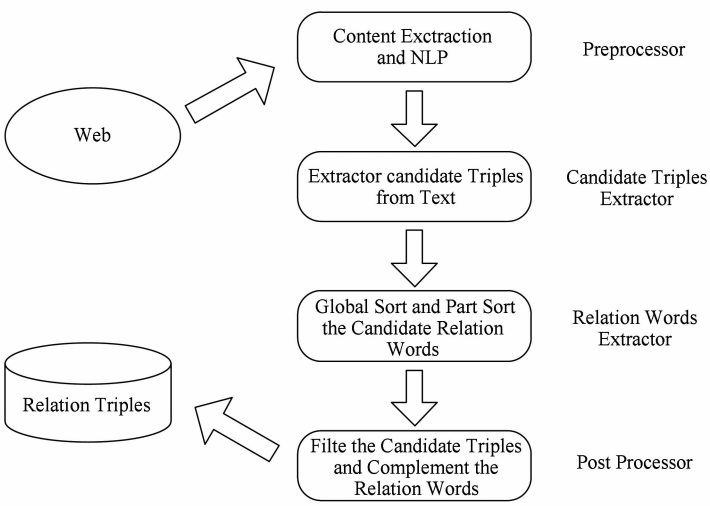


Fig. 1 Architecture of UnCORE.

图 1 面向大规模网络文本的开放式实体关系抽取模型

1.1 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 2 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. 使用哈尔滨工业大学语言技术平台(language technology platform, LTP)[8]对网页文本进行断句、分词、词性标注和命名实体识别.

1.2 生成候选三元组

为了更好地刻画关系三元组的抽取模型,我们

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

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2个生成候选关系三元组的限制条件.

1.2.1　实体之间的距离限制

图2中点(5,0.7457)表示2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实体关系.因此,在生成候选关系三元组的时候,规定2个实体之间词的数目最多不超过 $maxDistance$.

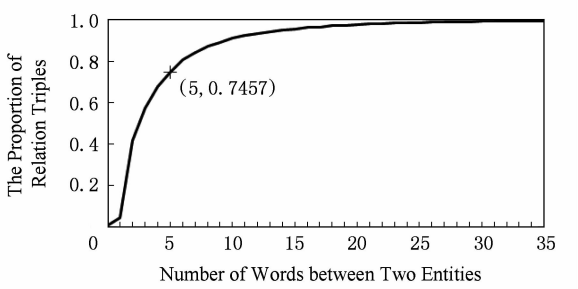


Fig. 2　The distribution of relation triples with different word distance.

图2　关系三元组数目在词距离上的分布情况

从图3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 $maxEntityDistance$.

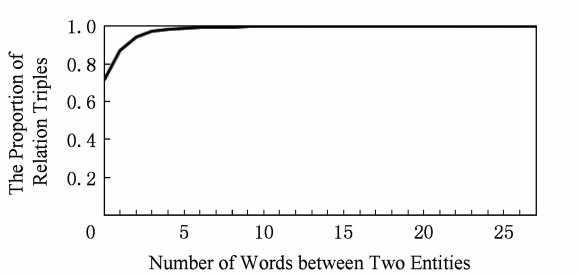


Fig. 3　The distribution of relation triples with different entity distance.

图3　关系三元组数目在实体距离上的分布

1.2.2　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.

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1个实体左边 $leftWordNumber$ 个名词和动词、第1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.

Table 1　The Distribution of Relation Triples with the Position of Relation Words

表1　指示词在句子中的位置分布

Position of Relation Words	Number of Triples	Proportion/%
Between Entities	3 177	75.36
Right of the Second Entity	609	14.44
Left of the First Entity	160	3.80
No Relation Words	240	6.40

1.3　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.

1.3.1　全局关系指示词排序

通过分析语料发现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,信息增益的计算如式(1)所示:

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

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$\}$,本文区分2个实体在句子中的先后顺序.

1.3.2　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.式(2)表述了关系指示词 rel 描述实体对类型 t 的实体关系的能力.

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

其中,$p(t|rel)$ 表示实体对类型 t 在关系指示词 rel 下的概率,$c(rel,t)$ 表示 rel 和 t 的共现的三元组数量.如果 rel 是关系指示词,那么 $p(t|rel)$ 保证了 rel 只能描述类型 t 的实体对之间的关系,$c(rel,t)$ 则保证了 rel 的描述能力具备统计意义.

1.3.3 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算法1所示:

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 $\{RelationWords(t)\mid t\in types\}$.

步骤1. 令集合 $IGCandidateRelationWords$ 为 $CandidateRelationWords$ 按照 $IG(rel)$ 值降序排序结果;

步骤2. 令集合 $IGList$ 为 $IGCandidateRelationWords$ 的前 N 个元素;

步骤3. 对集合 $types$ 中的每个元素 t;

步骤3.1. 令集合 $scoreCandidateRelationWords(t)$ 为 $CandidateRelationWords$ 按照 $score(rel,t)$ 值降序排序结果;

步骤3.2. 令集合 $scoreList(t)$ 为 $scoreCandidateRelationWords(t)$ 的前 K 个元素;

步骤3.3. 令集合 $RelationWords(t)$ 为 $scoreList(t)$ 和 $IGList$ 的交集;

步骤4. 返回集合 $\{RelationWords(t)\mid t\in types\}$.

1.4 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.

1.4.1 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"中抽出的候选三元组(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.

1.4.2 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1个实体在句子中的位置, j 是第2个实体在句子中的位置.2条噪声句式:

1) 关系指示词包含动词且第2个实体后面第1个词语是动词,其形式化描述为

$$hasV(relationWords)\wedge isV(posj+1)\Rightarrow isErrorTriple(i,relationWords,j).$$

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中第2个实体后面第1个词语是"的",其形式化描述为

$$\neg hasV(relationWords)\wedge isDE(posj+1)\Rightarrow isErrorTriple(i,relationWords,j).$$

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2种句式,那么三元组将被从候选集合中删除.

1.4.3 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.上述例子中,补全关系指示词之后的三元组是(王树国,担任校长,哈尔滨工业大学).

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.

2 实 验

2.1 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10 GB文本,包含以下3个来源:

1) 百度百科[①]160万个网页;

2) 新浪音乐新闻[②];

3) 搜狗新闻语料[③].

① http://baike.baidu.com/

② http://ent.sina.com.cn/music/roll.html

③ http://www.sogou.com/labs/dl/ca.html

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2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2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.

2.2　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 $N=6\ 000$, $K=5\ 000$, $maxDistance=5$, $maxEntityDistance=0$, $leftWordNumber=0$, $rightWordNumber=0$ 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.

Table 2　Top 20 Relation Words in Each Domain

表2　关系指示词样例

Type of Entities Pair	Relation Words
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.

Table 3　Samples of Relation Triples Extrction

表3　网络文本中抽取的关系三元组样例

Type of Entities Pair	Relation Triples	Sentence
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	20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.

Table 4　Performance of Relation Triples Extraction on the Web Data

表4　在网络文本上的关系三元组抽取结果

Type of Entities Pair	Number of Triples		Precision/%	
	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
Accuracy			68.01	80.97

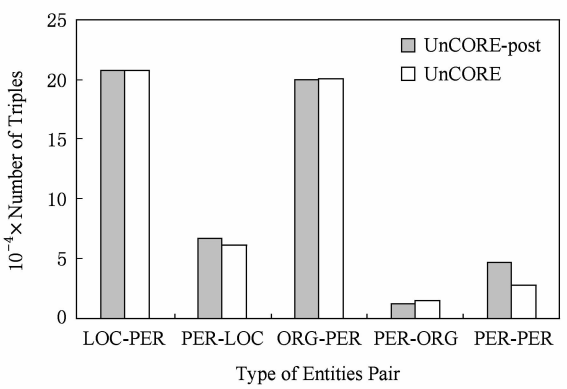


Fig. 4　The Number of Corret Relation Triples.

图4　正确的三元组数目

通过分析实验结果,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) 同时我们注意到 PER-LOC 的准确率较低,除了实体抽取错误以外,主要是由于关系指示词抽取不全造成,例如"李嘉诚前往陕西咸阳做公益,引村民围观."目前系统抽取出(李嘉诚,前往,陕西咸阳)这个三元组是不完整的.这类错误导致 PER-LOC 的效果比别的文本差很多.

5) 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 80%以上.

实体识别错误多,对实验结果影响较大,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).

Table 5 The Percentage of Triples with Wrong Entity

表 5 实体识别错误的三元组所占比例 %

Type of Entities Pair	Entity Error in all Triples	Entity Error in Wrong Triples
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

3 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 [C] //Proc of MUC-7. Stroudsburg, PA: ACL, 1998: 359-367

[2] Banko M, Cafarella M J, Soderland S, et al. Open information extraction from the Web [C] //Proc of IJCAI 2007. San Francisco: Morgan Kaufmann, 2007: 2670-2676

[3] Wu F, Weld D S. Open information extraction using Wikipedia [C] //Proc of ACL 2010. Stroudsburg, PA: ACL, 2010: 118-127

[4] Surdeanu M, Tibshirani J, Nallapati R, et al. Multi-instance multi-label learning for relation extraction [C] //Proc of the EMNLP 2012. Stroudsburg, PA: ACL, 2012: 455-465

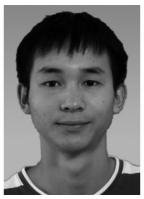
[5] Fader A, Soderland S, Etzioni O. Identifying relation for open information extraction [C] //Proc of the EMNLP 2011. Stroudsburg, PA: ACL, 2011: 1535-1545

[6] Yao L, Riedel S, McCallum A. Unsupervised relation discovery with sense disambiguation [C] //Proc of the 50th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2012: 712-720

[7] Wang Lifeng. Research on domain adaptive chinese entity relation extraction [D]. Harbin: Harbin Institute of Technology, 2011 (in Chinese)
(王莉峰. 领域自适应的中文实体关系抽取研究[D]. 哈尔滨: 哈尔滨工业大学, 2011)

[8] Che Wanxiang, Li Zhenghua, Liu Ting. LTP: A Chinese language technology platform [C] //Proc of the Coling 2010. Stroudsburg, PA: ACL, 2010: 13-16

Qin Bing, born in 1968. Professor and PhD supervisor in the School of Computer Science and Technology, Harbin Institute of Technology. Member of China Computer Federation. Her main research interests include Chinese information processing, information extraction and discourse analysis.



Liu An'an, born in 1989. Master in the School of Computer Science and Technology, Harbin Institute of Technology. His main research interests include entity relation extraction.

Liu Ting, born 1972. Professor and PhD supervisor. Senior member of China Computer Federation. His main research interests include natural language processing and social computing.

2015 年起《计算机研究与发展》双月将固定领域专题

致广大读者和作者:

本刊从 2015 年起将双数期约 1/2 版面固定为某个领域,每年将策划该领域的一个热点主题进行集中报道.具体的征文通知将在专题发表前 6 个月发布,请关注期刊网站!

此外,本刊依然欢迎自由来稿.谢谢!

具体领域分布及执行领域编委如下:

刊期	领域	领域编委	投稿方式	截稿日期
2 期	软件技术(含数据库)	孟小峰 xfmeng@ruc.edu.cn 中国人民大学	期刊网站投稿, 备注填写"年+专题名称"	上一年 10 月 1 日左右 (以具体征文通知为准)
4 期	网络技术	林闯 chlin@tsinghua.edu.cn 清华大学	期刊网站投稿, 备注填写"年+专题名称"	上一年 12 月 1 日左右 (以具体征文通知为准)
6 期	体系结构	刘志勇 zyliu@ict.ac.cn 中国科学院计算技术研究所	期刊网站投稿, 备注填写"年+专题名称"	当年 2 月 1 日左右 (以具体征文通知为准)
8 期	人工智能	周志华 zhouzh@nju.edu.cn 南京大学	期刊网站投稿, 备注填写"年+专题名称"	当年 4 月 1 日左右 (以具体征文通知为准)
10 期	信息安全	曹珍富 zfcao@sjtu.edu.cn 上海交通大学	期刊网站投稿, 备注填写"年+专题名称"	当年 6 月 1 日左右 (以具体征文通知为准)
12 期	应用技术	郑庆华 qhzheng@mail.xjtu.edu.cn 西安交通大学	期刊网站投稿, 备注填写"年+专题名称"	当年 8 月 1 日左右 (以具体征文通知为准)