DOI: 10.11992/tis.201707039
网络出版地址: http://kns.cnki.net/kcms/detail/23.1538.TP.20180409.1727.014.html

基于知识库的开放领域问答系统

张涛, 贾真, 李天瑞, 黄雁勇
（西南交通大学 信息科学与技术学院, 四川 成都 611756）

摘　要: 问答系统能够理解用户问题, 并直接返回答案。现有问答系统大多是面向领域的, 仅能回答特定领域的问题。文中提出了基于大规模知识库的开放领域问答系统实现方法。该系统首先采用自定义词典分词和 CRF 模型相结合的方法识别问句中的主体; 其次, 采用模糊匹配方法将问句中的主体与知识库中实体建立链接; 然后, 通过相似度计算以及规则匹配等多种方法识别问句中的谓词并与知识库实体的属性建立关联; 最后, 进行实体消歧和答案获取。该系统平均 F-Measure 值为 0.695 6, 表明所提方法在基于知识库的开放领域问答上具有可行性。

关键词: 问答系统; 开放领域; 实体识别; 实体链接; 知识库
中图分类号: TP391.1　　**文献标志码:** A　　**文章编号:** 1673−4785(2018)04−0557−07

中文引用格式: 张涛, 贾真, 李天瑞, 等. 基于知识库的开放领域问答系统[J]. 智能系统学报, 2018, 13(4): 557–563.
英文引用格式: ZHANG Tao, JIA Zhen, LI Tianrui, et al. Open-domain question-answering system based on large-scale knowledge base[J]. CAAI transactions on intelligent systems, 2018, 13(4): 557–563.

Open-domain question-answering system based on large-scale knowledge base

ZHANG Tao, JIA Zhen, LI Tianrui, HUANG Yanyong
(School of Information Science and Technology, Southwest Jiaotong University, Chengdu 611756, China)

Abstract: Question-answering (QA) systems can understand user questions and return answers directly. Currently, most QA systems can only answer questions pertaining to specific domains. In this paper, we propose a method for constructing an open-domain QA system based on a large-scale knowledge base. First, we present an approach based on a visual dictionary and a conditional random field (CRF) model to identify the subject in question. Next, we use a fuzzy matching method to link the entity in question to that in the knowledge base, and apply similarity computation and rule matching methods to recognize the question predicates and link them to the attributes of the knowledge entity. Lastly, we implement entity disambiguation and answer retrieval. The mean F-measure value of the proposed system is 0.695 6, which indicates the feasibility of the proposed method for an open-domain QA system for a large-scale knowledge base.
Keywords: question-answering system; open domain; entity recognition; entity linking; knowledge base

信息检索是目前互联网时代的一个热门研究方向。搜索引擎则是一种常见的信息检索手段, 它根据用户输入的查询语句进行关键字匹配, 并对结果进行排序, 返回包含关键字的页面链接, 并不会直接给出问题的答案, 需要用户自己浏览网页才能得到想要的答案。但是问答系统却能够克服这个缺点, 用户输入问句后问答系统会给出一个准确简洁的答案。

本文提出了一种基于知识库的开放领域自动问答系统。首先, 对问句进行实体识别, 即需要明确问句问的是关于哪个实体的, 实体识别是问答系统中非常重要的一个部分, 本文提出了自定义词典分词与 CRF 模型相结合的命名实体识别方法来完成该系统中的命名实体识别; 其次, 对实体进行链接, 即用问句中识别出的实体查询知

收稿日期: 2017−07−25.　**网络出版日期:** 2018−04−10.
基金项目: 国家自然科学基金项目 (61573292); 国家自然科学基金青年科学基金项目 (61603313).
通信作者: 张涛. E-mail: tzhangswjtu@163.com.

识库，返回实体名称相同或相近的实体信息，本文采用模糊匹配中后模糊匹配的方法来进行实体链接；然后，对谓词进行识别，即需要明确问句所问内容与实体的哪个属性相关，谓词识别同样也是该系统中一个非常重要的部分，本文采用了直接谓词匹配、词汇字面相似度和语义相似度及规则映射表 3 种方法来完成对问句中谓词的识别；最后，进行实体消歧、获取答案，由于知识库中存在很多同名实体，而且大多数同名实体的属性也相同或者相似，这就会导致答案会有多条，所以需要对包含答案的实体进行消歧，获取唯一的答案。

1　相关研究

问答系统 (question answering system, QA) 目前已经成为人工智能和自然语言处理领域中一个备受关注并具有广泛发展前景的研究方向[1]。现有的问答系统可以分为：1) 基于搜索引擎的问答系统；2) 基于社区的问答系统；3) 基于知识库的问答系统；4) 基于文本的问答系统。基于搜索引擎的问答系统首先需要解析问句，获得问句主体及类型，然后在搜索引擎返回的检索结果中按照问句类型抽取答案[2]。基于社区的问答主要是问句与互联网上的社区中用户提出的问句经过相似度计算返回结果[3]。而基于知识库的问答系统最主要的工作是进行问句理解[4-10]，提取出问句中的主体和谓词。比如 Poon[11]、Yahya[12]和 Berant[13-15]分别提出了基于语义分析的问答系统构建方法，其主要思路是先抽取问句中的主体和谓词，然后转化为 SPARQL 结构化查询语言，之后再与知识库交互得到答案。基于自由文本的问答系统则是从非结构化文本中抽取问句所问的答案。例如，Zheng[16]提出了一种从网页文本中抽取问句答案的方法。

2　开放领域问答系统方法概述

2.1　知识库预处理

1) 对 subject 中的英文大小写进行统一

因为知识库中的 subject 和问句中的 subject 存在英文大小写不统一的问题，所以将知识库 subject 中的英文都转化为小写。

2) 去除 subject 中的一些特殊的字符

在知识库 subject 中存在很多特殊字符比如“-”、“+”、“.”等。这些字符会影响分词和实体链接的结果，所以将 subject 中的这些特殊字符去掉。

2.2　实体识别

本文的实体识别采用的是自定义词典分词和 CRF 模型相结合的方法。实体识别的详细流程如图 1 所示。

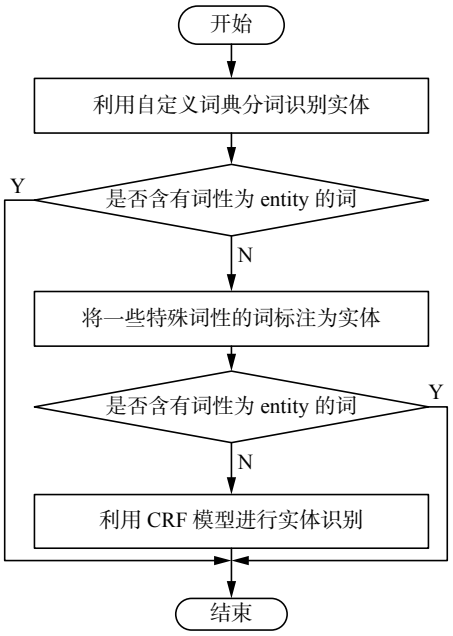

图 1　实体识别流程图

Fig. 1　Procedure of entity recognition

2.2.1　自定义词典分词识别实体

本文采用西南交通大学分词系统，它允许加载自定义分词词典进行分词。自定义词典分词过程中，首先是根据词典中的词进行组块分词，若在自定义词典中不存在再使用原分词算法进行分词和词性标注。由此本文将知识库中的 subject 提取出来构建自定义词典用于分词，以识别问句中的 subject。但是若将知识库中的所有 subject 都作为词典，可能问句中的普通词也被识别为实体。所以需要对 subject 进行筛选后再加入词典。本文将知识库中的 subject 进行分词，若 subject 分词之后词的数目大于 2，则将该 subject 加入词典。通过此方法构建的词典中包含 2 761 745 个实体词，其中部分词典如表 1 所示。

表 1　自定义分词词典

Table 1　The dictionary of custom segment

Subject	词性	Subject	词性
早安起床吻	entity	体外受精胚胎移植	entity
成都汇康医院	entity	我的知己在街头	entity
不倒翁的奇幻旅程	entity	鞍山巴黎花园	entity
成都五牛足球俱乐部	entity	字体转换器	entity
河北省滨海农业研究所	entity	幸福的拾荒者	entity
……	…	…	…

若问句中没有包含词典中的词，则利用分词系统进行实体识别。例如，若问句中出现词性为

nr、nh、nt 等的词, 则将其标注为实体。具体标记的词性和词性说明如表 2 所示。问句中没有包含这些词性, 则采用图 1 的算法进行实体识别。

表 2　词性和词性说明表

Table 2　The part of speech and its instruction

词性	词性说明	词性	词性说明
nr	人名	nt	机构团体名
ns	地名	nh	医药疾病等健康相关名词
nb	生物名	nf	食品
g	学术词汇		

2.2.2　基于 CRF 的实体识别

1) CRF 模型介绍

条件随机场算法 (conditional random field algorithm, CRF) 被认为是一种基于图模型的算法, 由于该图是一种无向图, 因此也是一种无向图模型[17]。在图模型中, 其中的结点表示算法的输入序列和输出序列, 主要用来计算输出序列的概率问题。条件随机场算法最主要的功能就是让无向图中的各个结点呈现出线性结构。线性的条件随机场算法同时也是一个有限的状态机, 该状态机可以进行线性数据的序列标注工作。当有限状态机接收到输入的序列后, 线性的条件随机场算法可以计算出输出序列, 并且能够返回该输出序列的条件概率。在条件随机场算法中, 训练模型的主要功能就是得到条件概率最大的模型[18]。

2) CRF 实体识别

由于建立词典和标记特殊词性的实体识别方法不能完全识别出所有问句中的实体, 所以本文采用了 CRF 算法来做进一步的实体识别。CRF 是一种常见的命名实体识别方法, 并取得了比较好的实验结果。用 CRF 做命名实体识别的步骤如下:

1) 人工标注一些问句中的实体作为训练数据;

2) 用标注好实体的训练数据训练 CRF 模型;

3) 用 CRF 模型来对未标注实体的问句进行实体标注。

本文利用已知实体回标的方式自动获取训练数据。即通过 3.2.1 中的方法得到实体所在的问句作为实体识别的训练数据。但是这些训练数据中存在少量错误, 需要人工再对训练数据做出修正。我们一共标注了约 13 000 条训练数据。

2.3　实体链接

命名实体链接是指将文本中已经识别出的命名实体链接到知识库中的一个具体真实实体的过程。实体链接是目前自然语言处理领域的一种常用技术。本文采用的是模糊匹配中后模糊匹配的方法进行实体链接。具体方法步骤如下:

1) 将问句中识别出的实体去除其中含括号的消歧项, 只保留实体原名称, 如: “武汉大学学报(医学版)/entity”, 去除括号内消歧项, 只保留“武汉大学学报”。

2) 用去除消歧项的实体名称到知识库中进行后模糊匹配查找, 后模糊匹配即忽略知识库中实体名称后面的所有词, 只关注知识库中 subject 的前 n 个字是否与所要查找的实体名称是否一致, n 为待查找实体名称的长度。如“武汉大学学报”, 只关注知识库 subject 的前 6 个字是否与“武汉大学学报”完全匹配。

3) 由于后模糊匹配有可能找出一些实体名称根本不可能是一个实体的 subject, 如在知识库中用后模糊匹配查找“李明”, 这会返回一下几个实体: “李明 (苏州大学教授)”、“李明慧 (大连轻工业学院副教授)”、“李明 (中医药研究教授)”等。其中“李明慧 (大连轻工业学院副教授)”就与“李明”不可能是同一个实体, 这是后模糊匹配忽略了“李明”之后的所有词才导致了这个结果, 对于这种情况本文首先对查找出来的实体去除括号里面的消歧项, 然后与原查找名称进行比较, 若不相等则删掉该实体。

2.4　谓词匹配

谓词匹配的作用是确定问句问的是实体的哪个属性 (谓词)。本文采用了多种谓词匹配方法, 其中包括直接谓词匹配法、词汇字面相似度和语义相似度相结合的方法、规则映射法以及同义词表映射法。

2.4.1　直接谓词匹配法

在自然语言问句中有经常会出现很多问句中直接包含谓词的情况, 可以通过直接谓词匹配的方法进行查找答案。具体方法如下: 提取所有链接实体的谓词作为候选谓词, 然后进行直接匹配。若问句中除去实体之外的词中包含谓词, 则将该谓词所对应的三元组加入三元组列表中。

2.4.2　词汇字面相似度和语义相似度相结合的方法

由于语言表达的多样性, 问句中的很多谓词并不能通过直接匹配法找到。针对那些谓词不是直接包含在问句中或谓词以同义词的形式出现在问句中的情况, 本文采用了字面相似度和语义相似度相结合的方法来识别谓词。字面相似度计算公式为

$$\mathrm{sim} = \frac{n}{N} \tag{1}$$

式中：N 为链接实体中的某个谓词长度，遍历谓词的每个字 w，若 w 在问句中出现，则 n=n+1(n 的初始值为零)。

本文使用的语义相似度计算是基于《知网》的相似度算法，该方法是由夏天在《汉语词语语义相似度计算研究》[19]一文中所提出的。《知网》是一个以汉语和英语的词语所代表的概念为描述对象，以揭示概念与概念之间以及概念所具有的属性之间的关系为基本内容的常识知识库。在《知网》中，词汇对应于若干概念，而概念是以义原为基础通过知识库描述语言进行定义的，即概念的义项表达式，义原又通过多种关系进行描述，如上下位关系等，目前大多数学者基于《知网》的词汇语义相似度计算，其思想是整体相似度可由部分相似度加权平均进行计算。

本文字面相似度和语义相似度结合匹配谓词的步骤如下：

1) 计算谓词和问句的字面相似度，若字面相似度等于 1，则表示谓词直接出现在问句中，并将该谓词所对应的三元组加入三元组列表中；

2) 计算谓词与问句中去除实体后的每个词的语义相似度，若存在与谓词语义相似度等于 1 的词，则表示谓词出现在问句中，并将该谓词所对应的三元组加入三元组列表中；

3) 若步骤 1)、2) 后，问句中都没有谓词，则取步骤 1) 中的字面相似度 sim 在 0.5~1 的谓词，计算这些谓词与问句中去除实体后的每个词的语义相似度，取其中之最大的谓词所对应的三元组加入三元组列表中。

2.4.3　规则映射法

由于 2.4.1 和 2.4.2 都是以词为单位进行谓词查找的，但是有些属性是多个词共同表达的，如问句“……什么时候去世”，谓词“去世地点”和“去世时间”都包含“去世”，但“什么时候”表达了时间。我们制定了规则解决这一类问句的谓词识别问题。例如，对于“什么时候”后接一个动词的，可将其映射为动词+(时间|日期)。如“什么时候去世”，“去世”是动词，什么时候表示时间，就可以映射为“去世时间”或“去世日期”。表 3 列出了规则示例。

表 3　规则映射表

Table 3　Examples of rule set

规则名称	谓词映射表
什么时候.*?/v	v+时间|日期
如何.*?/v	v+方式|方法|途径
在哪.*?/v	v+地点|地

若经过步骤 2.4.1 和 2.4.2 后，仍然还未找到谓词，则验证问句中是否包含表 3 中的规则，若包含则用规则映射后的词语去匹配链接实体在知识库中的每个谓词，匹配谓词的方法为步骤 2) 中的词汇字面相似度和语义相似度相结合的方法，若找到相同或相似的谓词则将该谓词所对应的三元组加入三元组列表中。

2.4.4　同义词表映射法

在常见问句中还有一些问句中的谓词是以同义词的形式存在，或者每个问句格式可以对应一些特定的谓词。如问句中存在“有多高”这种疑问词，则可以将谓词映射为“高度”。对于这些问句本文采用建立同义词表的方法来做谓词映射。如“什么意思”其同义词表可为“含义|意义|解释”。本文建立了同义词映射表，部分同义词表如表 4 所示。

表 4　同义词映射表

Table 4　Examples of synonym list

问句中的词汇	同义词
什么意思	含义|意义|解释
有多长	长度|片长|时间长度|路线长度
有多少人	人口|人数|全院人数
是哪的人	老家|籍贯|出生地
有几笔	笔画|笔划
…	…

同样，问句如果经过 2.4.1、2.4.2 和 2.4.3 这 3 个步骤后还没有找到相应的答案，则验证问句中是否包含表 4 中的词汇，若包含则用同义词表中的词语去匹配实体的每个谓词，若找到相同或相似的谓词则将该谓词所对应的三元组加入三元组列表中。

2.5　获取答案

在完成谓词匹配后就能够得到谓词所对应的三元组列表，三元组列表中每个对应 object 都是问题的答案。三元组列表包含一个或多个答案。这是因为知识库中存在很多同名实体，而且同名实体其谓词基本上是相同或相似的，导致有些问句就会出现一个或多个答案。需要从答案列表中选择一个作为最终的答案，即答案筛选。本文采用的方法是对所有答案对应的实体名称进行实体消歧，消歧后得到三元组中的 object 作为该问题的答案。

本文实体消歧的方法如下：

1) 取分词后问句中所有名词 (实体词除外)；

2) 对三元组列表中的 subject 进行分词并去除其中的标点符号;

3) 将所有 subject 分词后的所有词与步骤 1) 中的所有名词计算词汇语义相似度, 取其中语义相似度最大的 subject 所对应的 object 作为问句的答案。

如“西游记属于什么类型的书呢”, 完成实体识别和分词后的句子为“西游记/entity 属于/v 什么/ry 类型/n 的/ude1 书/n 呢/y”。该问句中, “西游记”被识别为实体词, 即主体 subject。经过上述步骤 1), 取问句中的名词 “类型/n 书/n”, 其他词过滤掉。而经过谓词识别后得到知识库三元组如表 5 所示。其中“西游记 (吴承恩等著作小说)”, 经过步骤 2) 的结果为“西游记/nz 吴承恩/nr 等/v 著/uzhe 作/v 小说/n”。然后将步骤 1) 和 2) 进行交叉计算语义相似度, 取其中最大的语义相似度作为该实体与问句的相似度。由于“小说”和“书”的语义相似度为 1, 所以该实体与问句的语义相似度为 1。由于其他 subject 与问句的语义相似没有比 1 大的, 所以该问句的答案为 subject 为“《西游记》(吴承恩等著作小说)”所对应的 object, 即“古典神魔小说”。

表 5　“西游记”实体链接和谓词匹配后的三元组列表

Table 5　Triple list of “Journey to the West” after entity link and predicate matching

subject	predicate	object
西游记	类型	魔幻, 剧情
西游记 (吴承恩等著作小说)	类别	古典神魔小说
西游记 (2010 年新版动画片)	类型	冒险, 搞笑, 神魔, 电视
西游记 (元代杨景贤创作杂剧)	类型	杂剧
西游记 (2010 年程力栋执导电视剧)	类型	古装神话剧
…	…	…

3　实验结果及分析

3.1　实验数据集描述

1) 知识库

本文中使用的知识库是 NLPCC2016 测评中发布的知识库[19]。知识库是以三元组的形式给出的, 三元组格式为⟨subject, predicate, object⟩, 总共有 43 063 796 条三元组, 其示例数据如表 6 所示。

2) 数据集

①训练数据集: 本文采用的训练数据集是 NLPCC2016 测评提供的数据[19], 其中包含 14 609 测试问句。

②测试数据集, 本文采用的测试数据集也是 NLPCC2016 测评提供的数据[19-20], 其中包括 9 870 条测试问句。

表 6　知识库三元组

Table 6　Triples in knowledge base

Subject	Predicate	Object
成都	别名	成都
成都	中文名称	成都
成都	外文名称	Chengdu
成都	行政区类别	地级市
成都	所属地区	中国西南
成都	邮政区码	610000
…	…	…

3.2　实验评价指标

此次评测使用了 F-Measure 值来衡量开放领域问答系统的性能。其相应的计算公式为

$$\text{Averaged}F_1 = \frac{1}{|Q|}\sum_{i=1}^{|Q|} F_i \tag{2}$$

当C_i中的值都不在A_i中, 则F_i为 0, 否则, 使用式 (3) 计算F_i的值:

$$F_i = \frac{2\times\dfrac{\#(C_i,A_i)}{|C_i|}\times\dfrac{\#(C_i,A_i)}{|A_i|}}{\dfrac{\#(C_i,A_i)}{|C_i|}+\dfrac{\#(C_i,A_i)}{|A_i|}} \tag{3}$$

式中: $\#(C_i,A_i)$为C_i和A_i中拥有相同答案的个数, $|C_i|$和$|A_i|$是各自中答案的总数。

3.3　实验评测结果

本文根据谓词识别方法的不同做了几组实验, 分别为直接匹配、直接匹配+组合谓词、直接匹配+词语相似度、以及所有方法组合即本文最终采用的方法, 其中词语相似度包含词语字面相似度以及词语语义相似度。实验结果如表 7 所示, 可见本文最终采用的方法的 F 值比其他几种方法的 F 值都要好。

表 7　评测结果

Table 7　Evaluation results in the task

System	F_1 Score
直接匹配	0.462 5
直接匹配+组合谓词	0.540 2
直接匹配+词语相似度	0.662 3
本文方法	0.695 6

4 结束语

针对 NLPCC2016 开放领域问答测评提供的问句，本文提出了一种基于知识库的开放领域问题的问答系统构建方法，该方法首先对问句进行实体识别，即提取出问句中的实体，在识别出问句实体之后则进行谓词匹配，将谓词匹配度高的 object 作为问句的答案返回。为了提高问答系统的准确率，本文还对知识库进行了适当清理。实验表明基于本文所提出的方法的平均 F 值为 0.695 6，充分证明了本文所提出方法的可行性。

参考文献：

[1] MOONEY R J. Learning for semantic parsing[C]//Proceedings of the 8th International Conference on Computational Linguistics and Intelligent Text Processing. Berlin, Heidelberg: Springer, 2007: 311–324.

[2] FILMAN R E, PANT S. Searching the internet[J]. IEEE internet computing, 1998, 2(4): 21–23.

[3] JEON J, CROFT W B, LEE J H. Finding similar questions in large question and answer archives[C]//Proceedings of the 14th ACM International Conference on Information and Knowledge Management. New York, NY, USA: ACM, 2005: 84–90.

[4] ZETTLEMOYER L S, COLLINS M. Learning to map sentences to logical form: structured classification with probabilistic categorial grammars[C]//Proceedings of the 21th Conference on Uncertainty in Artificial Intelligence. Arlington, Virginia, USA: AUAI, 2005: 658–666.

[5] WONG Y W, MOONEY R J. Learning for semantic parsing with statistical machine translation[C]//Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics. Stroudsburg, PA, USA: The Association for Computational Linguistics, 2006: 439–446.

[6] WONG Y W, MOONEY R J. Generation by inverting a semantic parser that uses statistical machine translation [C]//Proceeding of the Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics. Rochester, New York, USA: The Association for Computational Linguistics, 2007: 172–179.

[7] ZETTLEMOYER L S, COLLINS M. Online learning of relaxed CCG grammars for parsing to logical form[C]// Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Prague, Czech Republic: The Association for Computational Linguistics, 2007: 678–687.

[8] KWIATKOWSKI T, ZETTLEMOYER L, GOLDWATER S, et al. Lexical generalization in CCG grammar induction for semantic parsing[C]//Proceedings of 2011 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA, USA: The Association for Computational Linguistics, 2011: 1512–1523.

[9] KWIATKOWSKI T, ZETTLEMOYER L, GOLDWATER S, et al. Inducing probabilistic CCG grammars from logical form with higher-order unification[C]//Proceedings of 2010 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA, USA: The Association for Computational Linguistics, 2010: 1223–1233.

[10] YE Zhonglin, JIA Zheng, YANG Yan, et al. Research on open domain question answering system[C]//LI Juanzi, JI Heng, ZHAO Dongyan, et al. Proceedings of the 4th International Conference on Natural Language Processing and Chinese Computing (NLPCC2015). Cham, Germany: Springer, 2015: 527–540.

[11] POON H, DOMINGOS P. Unsupervised semantic parsing[C]//Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA, USA: The Association for Computational Linguistics, 2009: 1–10.

[12] YAHYA M, BERBERICH K, ELBASSUONI S, et al. Natural language questions for the web of data[C]//Proceedings of 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Stroudsburg, PA, USA: ACL, 2012: 379–390.

[13] YAO Xuchen, BERANT J, VAN DURME B. Freebase QA: information extraction or semantic parsing[C]//Proceedings of the ACL 2014 Workshop on Semantic Parsing. Baltimore, Maryland USA: The Association for Computational Linguistics, 2014: 82–86.

[14] BERANT J, CHOU A, FROSTIG R, et al. Semantic parsing on freebase from question-answer pairs[C]//Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA, USA: Association for Computational Linguistics, 2013: 1533–1544.

[15] BERANT J, LIANG P. Semantic parsing via paraphrasing[C]//Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. Baltimore, Maryland, USA: Association for Computational Linguistics, 2014: 1415–1425.

[16] ZHENG Zhiping. AnswerBus question answering

system[C]//Proceedings of the Second International Conference on Human Language Technology Research. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc, 2002: 399–404.

[17] LIU F Y, LIN G S, SHEN C H. CRF learning with CNN features for image segmentation[J]. Pattern recognition, 2015, 48(10): 2983–2992.

[18] GANAPATHY S, VIJAYAKUMAR P, YOGESH P, et al. An intelligent CRF based feature selection for effective intrusion detection[J]. International Arab journal of information technology, 2016, 13(1): 44–50.

[19] 夏天. 汉语词语语义相似度计算研究[J]. 计算机工程, 2007, 33(6): 191–194.
XIA Tian. Study on Chinese words semantic similarity computation[J]. Computer engineering, 2007, 33(6): 191–194.

[20] WU Yunfang, LI Wei. Overview of the NLPCC-ICCPOL 2016 shared task: Chinese word similarity measurement[C]// Natural Language Understanding and Intelligent Applications. Cham, Germany: Springer, 2016: 828–839.

作者简介:

张涛, 男, 1989 年生, 硕士研究生, 主要研究方向为中文信息处理、信息抽取、智能问答。

贾真, 女, 1975 年生, 讲师, 博士, 主要研究方向为自然语言理解、中文信息处理、信息抽取、大数据。

李天瑞, 男, 1969 年生, 教授, 博士生导师, 博士, 主要研究方向为智能信息处理、数据挖掘、云计算和大数据。