

文章编号: 1003-0077(2019)05-0024-07

基于桥连接的词典学习方法的语义解析

陈波, 孙乐, 韩先培

(中国科学院 软件研究所 中文信息处理实验室, 北京 100190)

摘要: 现阶段的语义解析方法大部分都基于组合语义, 这类方法的核心就是词典。词典是词汇的集合, 词汇定义了自然语言句子中词语到知识库本体中谓词的映射。语义解析一直面临着词典中词汇覆盖度不够的问题。针对此问题, 该文在现有工作的基础上, 提出了基于桥连接的词典学习方法, 该方法能够在训练中自动引入新的词汇并加以学习, 为了进一步提高新学习到的词汇的准确度, 该文设计了新的词语—二元谓词的特征模板, 并使用基于投票机制的核心词典获取方法。该文在两个公开数据集(WebQuestions 和 Free917)上进行了对比实验, 实验结果表明, 该文方法能够学习到新的词汇, 提高词汇的覆盖度, 进而提升语义解析系统的性能, 特别是召回率。

关键词: 语义解析; 词典学习; 组合语义; 覆盖度

中图分类号: TP391 **文献标识码:** A

A Bridge-based Lexicon Learning Method for Semantic Parsing

CHEN Bo, SUN Le, HAN Xianpei

(Chinese Information Processing Laboratory, Institute of Software,
Chinese Academy of Sciences, Beijing, 100190, China)

Abstract: Current Semantic Parsers are mainly based on compositional semantics, with a strong dependent on lexicon. Lexicon is a set of vocabularies, which define the mappings between words or phrases from natural language sentences and predicates from knowledge base ontology. In order to deal with the low-coverage of lexicon, this paper proposes a bridge based lexicon learning method for semantic parsing on the basis of existing work. This method can bring in new vocabularies during training and learn a new lexicon with high-coverage. Furthermore, this paper designs a new word-predicate feature template and utilizes voting to gain core vocabularies for more accurate lexicon. Experiments results on two benchmarks: WebQuestions and Free917, show that our method can learn new vocabularies for improving the coverage of lexicon, with a side-effect on parsing improvement.

Keywords: semantic parsing; lexicon learning; compositional semantics; coverage

0 引言

语义解析是把自然语言句子解析成完整的、计算机可识别的、可计算的或可执行的语义表示。通常使用逻辑形式来表示语义, 如 Lambda—表达式。语义解析是实现自然语言理解的核心技术, 也是实现人机交互应用^[1-2]、智能机器人^[3]所必需的技术, 还对自然语言处理的其他领域(如机器翻译、问答、知识抽取等)具有促进作用。

现阶段的语义解析技术大都基于组合语义, 而

组合语义基于一条基本原则: 句子的意思是句子中各个成分的意思的组合。通常, 基于组合语义的语义解析系统都包含两个重要的模块: 词典和文法, 词典对应上述原则中各个成分的意思, 而文法对应成分的意思如何组合。词典和文法是基于组合语义的语义解析系统的核心和基础。文法一般是从语言现象中提炼出来的若干条组合规则。而词典通常都是需要学习的。随着语义解析技术逐渐面向开放域, 目标知识库本体中的谓词(predicate)数目也越来越多, 而自然语言本身表达具有丰富性: 每一个意思可以由多个不同的词语来表达, 这使得词典学

收稿日期: 2018-02-16 定稿日期: 2018-03-22

基金项目: 国家自然科学基金(61433015, 61572477)

习越来越困难,其中词典中词汇覆盖度问题尤其显著。如何加强词典学习,提高词典中词汇覆盖度成了语义解析领域的一大研究难点和重点。

词典学习一方面是语义解析任务中不可或缺的一部分,另一方面它本身也可作为自然语言处理中一项十分重要的任务,词典可用于机器翻译、信息抽取等。近年来,语义解析领域开始有很多工作着手于词典学习,Cai^[4]等提出了基于模板的词典学习方法,该方法利用文本语料和知识库中实体对的共现,以及特定模板的约束来学习词语到谓词之间的映射,这种方法使用远距离(distant)监督学习算法能够学习到新的词汇来扩充词典,一定程度上提高了词典的覆盖度。但由于谓词数目实在太多,有很多谓词都无法依靠这种方法得到准确的词语与之匹配。Berant^[5]等在此基础上提出桥连接的方法,该方法在没有词语触发二元谓词的情况下,也就是词典中的词汇没有覆盖到的情况下,能够利用当前的解析结果信息(如已识别实体)和知识库本体的约束,自动引入二元谓词。该方法虽然没有学习新的词汇,但是桥连接充当了词汇的功能,在一定程度上能够纾解词典覆盖度不够的问题。然而,桥连接方法并没有学习到真正的词汇。

针对语义解析中词典中词汇覆盖度不够的问题,本文在现有工作的基础上提出了基于桥连接的词典学习方法。该方法在桥连接方法的基础上,在训练的过程中,每自动引入一个二元谓词,我们就从句子中抽取非实体性的内容词语(疑问词、形容词、名词和动词),并与该二元谓词配对,作为新的词汇加入到模型中。随着新词汇的大量加入,词典中会存在很多带“歧义”的词汇,同时也存在很多无用的甚至错误的词汇。为了让模型选择更好的词汇,我们设计了新的词语—二元谓词特征模板,该特征模板能够更好地捕捉词语和二元谓词之间的映射关系。为了截取无用的、错误的词汇,我们设计了基于投票机制的核心词汇获取方法,核心词汇虽然牺牲了一点语义解析系统的召回率,但是能够大大提升语义解析的效率,还能提高语义解析的准确率。

本文的主要工作总结如下:

(1) 针对语义解析任务中词典覆盖度不够的问题,我们在现有工作的基础上提出了基于桥连接的词典学习方法。

(2) 我们设计了新的词语—二元谓词的模板特征,来消除新词典中容易出现歧义的问题。

(3) 我们设计了基于投票机制的核心词汇获取方法来纾解新词典中存在无用词汇或者错误词汇所带来的影响。

1 相关研究工作及背景知识

1.1 语义解析

一直以来语义解析任务都吸引了大批学者,也涌现了很多工作。从方法层面这些工作大致可以划分为如下几类:①基于传统组合语义的方法;②基于语义图的方法;③基于序列到序列的方法。基于传统组合语义的语义解析系统^[6]一般由5个模块组成:文法、分析器、模型、学习器和执行器。其中文法包括词典和组合规则。现阶段常用的文法有组合范畴文法(CCG)^[6-7]和基于依存的组合语义文法(DCS)^[8]。基于语义图的方法^[9-10]借用语义图来表示句子的语义,一般依靠句法树、依存树、特定模板或者启发式算法来获取语义图。基于序列到序列的方法^[11-12]首先将逻辑表达式序列化,然后把语义解析问题转化为自然语言句子词语序列到逻辑表达式序列的翻译问题,并且可以使用递归神经网络模型来建模序列的生成。

1.2 词典及词典学习

词典是基于组合语义的语义解析方法的核心组件。词典是词汇的集合,每一个词汇表示一个自然语言中的词语到知识本体中谓词的映射。词汇一般分为三类:实体词汇、类别词汇和关系词汇。实体词汇对应知识本体中的实体;类别词汇对应知识本体中的一元谓词;关系词汇对应知识本体中的二元谓词。这三类词汇的实例分别对应图1中的词汇2、词汇1和词汇3。词典也是基于组合语义的语义解析方法的基础。图1展示了一个语义解析的例子,从例子中我们可以看到,词汇作用于解析过程中的最底层,可以说没有词汇,就无法自底向上地解析。由此引申出来词典学习这一项重要的任务。词典学习就是学习词汇,学习映射关系,一般实体词汇直接利用实体链接技术来获取,需要学习的是类别词汇和关系词汇,其中关系词汇尤其重要,也是最难的部分。

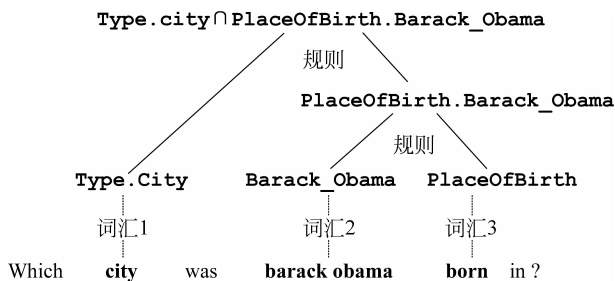


图1 基于组合语义的语义解析的例子

2 基准语义解析系统

这一节介绍我们用于对比系统的基准系统,我们采用基于传统组合语义的语义解析系统作为基准系统。具体的,我们的系统基于 SEMPRE^[5] 系统而搭建。该系统总共由 5 个模块^[13] 组成: 文法、模型、解析器、学习器和执行器(如图 2 所示)。其中文法模块是该系统的基础,也是核心,该模块包括词典和组合法;解析器模块负责利用文法对句子进行解析;模型模块负责对解析器所得到的候选解析结果进行打分排序;学习器模块负责利用训练数据对模型模块中的模型参数进行学习;执行器模块负责利用知识库对得到的逻辑表达式进行求值,得到句子所对应的答案。

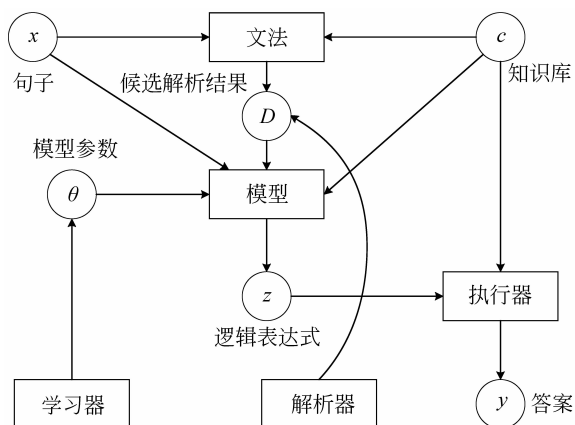


图2 基于组合语义的语义解析方法的基本框架图

2.1 桥连接

Berant^[5]等提出了桥连接方法来充当词汇的功能。具体的,当句子中的词语没有触发任何谓词的时候,该方法可以根据当前的解析结果自动引入二元谓词,如图 3 例子所示,二元谓词 Currency(通用货币)并不是由关系词汇触发而来,而是利用当前所得到的实体 SriLanka(斯里兰卡,一个国家)自动引

入的(所有与这个实体关联的二元谓词都会引入,模型再利用特征来选择最合适的)。该方法也是用来解决词汇覆盖度不够的问题。

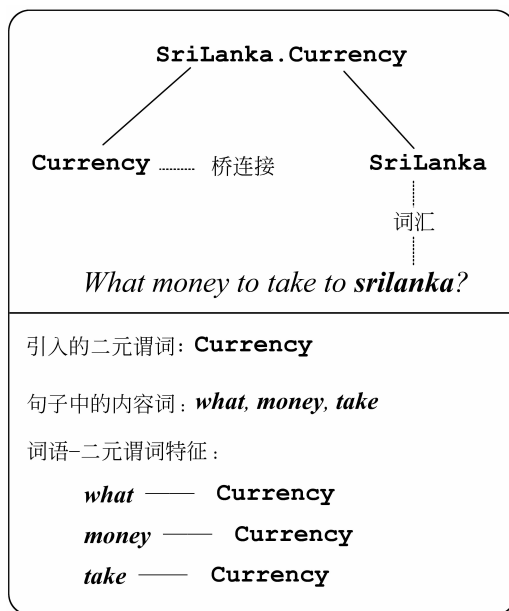


图3 桥连接方法示例

2.2 针对桥连接的词语——二元谓词的特征模板

Berant^[13]等针对桥连接设计了一个词语—二元谓词的特征模板(lemmaAndBridge),该特征模板在使用桥连接的时候激活,该特征模板用来描述句子中非实体性内容词(疑问词、动词、形容词和名词)与桥连接所引入的二元谓词之间的映射关系。桥连接与该特征模板的联合使用可以说是充当了隐式词汇的功能。

3 基于桥连接的词典学习的语义解析

这一节介绍我们提出的基于桥连接的词典学习的语义解析方法,首先我们介绍基于桥连接的词典学习方法,然后再介绍我们所采用的用于提高词典准确度的两个策略:新的词语—二元谓词的特征模板和基于投票机制的核心词汇获取。图 4 展示了我们方法的框架图。

3.1 基于桥连接的词典学习方法

为了解决词汇覆盖度不够的问题,我们在现有工作的基础上,提出基于桥连接的词典学习方法。Berant^[5]等提出的桥连接方法,能自动引入二元谓词,但他们的方法并没有学习到新的词汇,只是利用

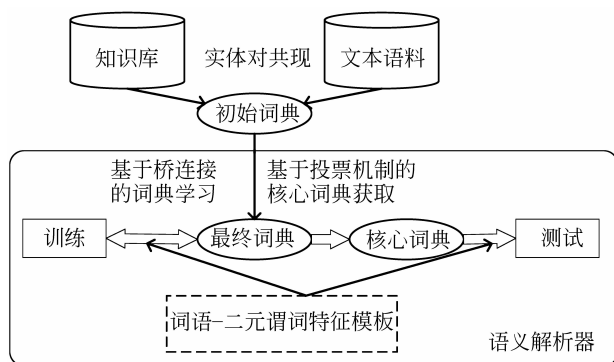


图4 我们方法的框架图

桥连接来充当词汇的角色。我们的方法在他们方法的基础上加以改进,可以引入新的词汇,并加以学习。

具体的,在解析的过程中,当桥连接被激活的时候,会有一个二元谓词被自动引入,而这个二元谓词并不是由句子中的某个词语触发的。实际上,这个二元谓词本应该由句子中的某个词语触发,也就意味着应该存在一个类别词汇,这个词汇是句子中某个词语到这个二元谓词的映射(图3中的例子,本应该存在 *money::Currency* 这个词汇),而正因为词典的覆盖度不够,词典缺失了很多词汇,使得在解析的过程中,不能够利用词汇触发来引入谓词。为此,我们在这个过程中引入新的词汇,并把新的词汇加入到当前的词典中。当一个桥连接激活的时候,我们从句子中抽取非实体性的内容词语(疑问词、动词、名词和形容词),把这些词语分别与引入的二元谓词配对,形成若干新的词汇,图3的例子中会加入3个新词汇,分别是:

what :: Currency
money :: Currency
take :: Currency

这些词汇中肯定包含了本应该存在的能够触发该二元谓词的词汇。我们把这些词汇加到现有词汇中,通过一定的样例训练以后,新加入的词汇中那些有效的词汇就会被模型通过特征“挑选”出来,并可用于后面的对其他例子的解析。

词典中加入了新的词汇,并且基于桥连接加入的词汇中只有一部分是准确的,而其他部分是不准确的,这部分不准确的词汇不仅影响语义解析系统的准确性,还影响其效率。为此,我们设计了新的词语—二元谓词的特征模板,还设计了基于投票机制的核心词汇获取方法,用来提高词典的准确率和解析系统的效率。

3.2 新的词语——二元谓词的特征模板

Berant^[14]等设计的词语—二元谓词的特征模板仅在通过桥连接引入二元谓词的情况下激活,实际上,这个特征模板对所有的二元谓词都是十分有效的,不管这个二元谓词是通过关系词汇引入的,还是通过桥连接方法引入的。我们的方法在通过桥连接方法学习到新词汇以后,在解析的时候将不再使用桥连接方法,具体的,第一轮训练的时候启用桥连接方法,用于引入新词汇,之后的训练和测试,都不再启用桥连接方法,这样就可以测试我们新学到的词汇的效果。为了让这个特征模板发挥更大的功效,我们设计了可用于所有二元谓词的词语—二元谓词的特征模板(*lemmaAndBinary*),该模板只要有二元谓词引入就会激活。

3.3 基于投票机制的核心词典获取

由于我们现在使用的词典很大,特别是采用了基于桥连接的词汇引入方法之后,词典中的很多词汇都是错误的,这给语义解析带来很大的困扰,不仅大大影响解析的效率,也影响解析的性能。为此,我们提出了基于投票机制的核心词汇获取方法,该方法旨在获取当前词典中有用的、准确度高的核心词汇。Artzi^[15]等也提出了类似的方法,用来压缩词典。

具体的,在训练的时候,对于每一个样例,模型利用当前的词典对此样例进行解析。对于在正确的解析中所使用到的类别词汇和关系词汇,我们做记一票处理,由于模型可以利用这些词汇得到准确的解析结果,我们认为这些词汇是重要的且准确的词汇。经过若干轮的训练,词汇票数越多,代表词汇越重要、越准确。为了抽取核心词汇,又为了保障每一个谓词都存在它的核心词汇,首先我们从词典中抽取票数超过1的词汇,然后对这些词汇以谓词为基准进行分组,并按照投票数进行排序,最后我们对每一个谓词都抽取投票数, *top K* 的词汇用来组成我们的核心词典。在这里我们没有使用概率的因素,而是直接利用硬性条件来进行筛选。我们认为一个词汇的投票数超过1就表明该词汇在某个例子的解析中充当了正确的角色,应该纳入到核心词典的考虑范围;之所以为每一个谓词选择 *top K* (实验中 $K=5$) 的词汇,是因为我们认为有 $K(K=5)$ 个不同的词语来表达一个谓词的意思,就算自然语言表达具有丰富性的特点,也已经具有一定的覆盖度了。

测试的时候,模型就使用核心词典来对测试样例进行解析。

4 实验

4.1 数据集

我们在两个公开数据集上都进行了对比实验: WebQuestions 和 Free917。

WebQuestions 数据集是由 Berant^[5] 等收集而来。该数据集总共有 5 800 个样例,每个样例由问句及其答案组成。按照标准,3 778 个样例用于训练,2 032 个样例用于测试。

Free917 数据集由 Cai^[4] 等收集而来。该数据集总共有 917 个样例,每个样例由问句及其逻辑表达式组成。按照标准,641 个样例用于训练,276 个样例用于测试。

两个数据集都使用 Freebase 作为知识本体。

4.2 实验设定

在两个数据集上,我们都采用标准的训练/测试数据分割,若进行开发测试,则将训练数据按 0.8/0.2 的比例切分,其中 80% 的数据用来训练,剩下 20% 的数据用来测试。系统评价指标同样采用公认的标准^[5]: 平均准确率。我们的实验都是基于 SEMPRES 系统完成。我们所使用的解析算法(除了桥连接部分)、模型的其他参数和学习算法都与基准系统相同。训练的迭代次数分别是 3 和 6。在第一次迭代训练的时候开启桥连接功能,在其他训练轮次和测试环节关闭桥连接功能。用于获取核心词汇的阈值 K 都设置为 5。

4.3 实验结果

4.3.1 对比不同的特征模板(A 组实验)

首先,我们来对比我们提出的新的特征模板 lemmaAndBinary(记为 FT_2) 和 Berant^[14] 等提出的特征模板 lemmaAndBridge(记为 FT_1)。由于 lemmaAndBridge 仅在使用桥连接的情况下生效,所以我们在基准系统上面验证两个特征模板的效果(都开启桥连接功能)。我们都在开发集上进行测试,在两个数据集上的结果如表 1 和表 2 所示。从结果中我们可以看到,特征模板 FT_1 和 FT_2 都非常有效, F_1 值都提升了 4%~6%, FT_2 比 FT_1 更加有效。

实际上, FT_2 特征模板是包含 FT_1 特征模板的。在后面的实验中我们都在系统中使用 lemmaAndBinary 特征模板。

表 1 在 WebQuestions 数据集上的 A 组实验

	P	R	$F_1(P)$
都不使用 FT_1 、 FT_2	41.6	45.4	39.5
使用 FT_1 , 不使用 FT_2	48.0	51.3	44.7
使用 FT_2 , 不使用 FT_1	46.3	52.4	45.5

表 2 在 Free917 数据集上的 A 组实验

	P	R	$F_1(P)$
都不使用 FT_1 、 FT_2	61.9	61.9	61.3
使用 FT_1 , 不使用 FT_2	65.1	66.0	65.1
使用 FT_2 , 不使用 FT_1	66.6	66.0	65.9

4.3.2 检验获取核心词典的作用(B 组实验)

为了验证获取核心词典的作用,我们对比了不获取核心词典而直接使用整个新词典的系统和使用获取到的核心词典的系统。我们都在开发集上进行测试,实验中都使用 lemmaAndBinary 特征模板,关闭桥连接功能,此外我们还在原始词典的基础上测试了获取核心词典的效果,在两个数据集上的结果如表 3 和表 4 所示。从结果中我们可以看到,基于投票机制的核心词典获取方法可以提高系统的准确率,虽然会牺牲一点召回率,但是对系统整体性能起到提高的作用。另外我们还发现获取核心词典的方法能够大大压缩词典的大小。系统所利用的核心词汇是很少的,只占用了原始词汇的 1/500 不到(在 WebQuestions 数据集,原始词汇中类别词汇和关系词汇的数量分别是 282 005 和 132 111,而核心词汇的数量分别是 66 和 335)。在后面的实验中,我们都在系统中使用基于投票机制的核心词典获取方法。

表 3 在 WebQuestions 数据集上的 B 组实验

	P	R	$F_1(P)$
原始词典(L_0)	48.6	57.5	48.6
基于 L_0 的核心词典	49.2	54.8	48.5
新词典(L_n)	50.1	58.0	49.7
基于 L_n 的核心词典	51.5	56.8	50.4

表 4 在 Free917 数据集上的 B 组实验

	P	R	$F_1(P)$
原始词典(L_o)	63.9	63.9	63.3
基于 L_o 的核心词典	64.5	62.8	64.1
新词典(L_n)	66.9	69.2	66.4
基于 L_n 的核心词典	67.5	68.3	67.0

4.3.3 主要结果(C 组实验)

我们在最终的系统中使用了新的特征模板 lemmaAndBinary 和获取核心词汇的策略,并与基准系统在测试集上进行对比实验,在基准系统中,我们开启了桥连接功能,并选择使用了 lemmaAnd-Bridge 特征模板,在我们的系统中,我们使用了 lemmaAndBinary 特征,并同时使用基于投票机制的核心词典获取方法,设置实验再次测试这两种策略带来的效果。在两个数据集上的结果如表 5 和表 6 所示。从结果中我们可以看到,基于桥连接的词典学习方法在使用 lemmaAndBinary 特征模板和基于投票机制的核心词典获取的情况下,比桥连接方法更有效。更重要的是,我们的方法能够学习到新的词汇,可提供后续分析和使用。表 7 展示了几个我们新学习到的词汇。这些词汇都是原始词典中不存在的词汇,而是在训练过程中,基于桥连接方法,自动引入的词汇,并加以学习而得来的,我们可以看到这些词汇也符合我们的直觉,是我们所需要的,有了这些新词汇,能够有效提高语义解析器的性能。

表 5 在 WebQuestions 数据集上的 C 组实验

	P	R	$F_1(P)$
基准系统	49.7	58.2	49.6
我们的系统	51.3	59.3	50.9
我们的系统- FT_2	48.9	55.5	46.7
我们的系统-核心词典	50.2	59.8	50.0

表 6 在 Free917 数据集上的 C 组实验

	P	R	$F_1(P)$
基准系统	65.1	66.0	65.1
我们的系统	68.3	69.0	67.6
我们的系统- FT_2	65.4	63.3	64.6
我们的系统-核心词典	66.8	70.3	66.0

表 7 新学习到的几个词汇

词语	谓词
college	Education
wife	Spouse
kill	Cause_of_Death
win	Champion

5 结论与展望

本文提出了基于桥连接的词典学习的语义解析方法。该方法利用桥连接方法自动引入新的词汇,并加以学习,为了提高词汇的准确度,本文还设定了新的词语—二元谓词的模板。为了压缩词典的大小,获取到核心词典,进一步提高词汇的有效性,本文还提出了基于投票机制的核心词典获取方法。实验结果表明,我们的方法相比基准系统有很大的提升。

目前,我们的词汇获取方法还仅仅局限于利用训练语料,而训练语料是有限的,也就意味着还有很大的局限性,也导致词典覆盖度问题不能彻底地得以解决。下一步我们将利用大量文本资料来扩充词典,进一步解决词典覆盖度不够的问题。

参考文献

- [1] 周俊生,曲维光,许菊红,等. 基于语义解析的中文 GIS 自然语言接口实现研究[J]. 中文信息学报,2014,28(06): 62-69,106.
- [2] 谢炯坤. 面向人机互动的自然语言理解的研究[D]. 合肥:中国科学技术大学博士学位论文,2015.
- [3] 卢栋才. 服务机器人任务理解的关键技术研究[D]. 合肥:中国科学技术大学博士学位论文,2017.
- [4] Qingqing Cai, Alexander Yates. Large-scale semantic parsing via schema matching and lexicon extension [C]//Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics. 2013: 423-433.
- [5] Jonathan Berant, Andrew Chou, Roy Frostig, et al. Semantic Parsing on Freebase from Question-Answer Pairs[C]//Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. 2013: 1533-1544.
- [6] Luke S Zettlemoyer, Michael Collins. Learning to map sentences to logical form: Structured classification

- with probabilistic categorial grammars[C]//Proceedings of the 21st Conference on Uncertainty in Artificial Intelligence. 2005: 658-666.
- [7] 朱耀邦. 基于 CCG 的语义解析算法改进与组合方法研究[D]. 南京:南京师范大学硕士学位论文,2015.
- [8] Percy Liang, Michael Jordan, Dan Klein. Learning dependency-based compositional semantics[J]. Computational Linguistics. 2013,39(2): 389-446.
- [9] Siva Reddy, Mirella Lapata, Mark Steedman. Large-scale semantic parsing without question-answer pairs [J]. Transactions of the Association for Computational Linguistics,2014,2(10): 377-392.
- [10] Wen-tau Yih, Ming-Wei Chang, Xiaodong He, et al. Semantic parsing via staged query graph generation: Question answering with knowledge base[C]//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics. 2015: 1321-1331.
- [11] Li Dong, Mirella Lapata. Language to logical form with neural attention[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. 2016: 33-43.
- [12] Robin Jia, Percy Liang. Data recombination for neural semantic parsing[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. 2016: 12-22.
- [13] Percy Liang. Learning executable semantic parsers for natural language understanding[J]. Communications of the ACM,2016,59(9): 68-76.
- [14] Jonathan Berant, Percy Liang. Imitation Learning of Agenda-based Semantic Parsers[J]. Transactions of the Association for Computational Linguistics,2015, 3: 545-558.
- [15] Yoav Artzi, Dipanjan Das, Slav Petrov. Learning compact lexicons for ccg semantic parsing[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. 2014: 1273-1283.



陈波(1989—),通信作者,博士,主要研究领域为自然语言处理、语义解析。
E-mail: chenbo@nfs.iscas.ac.cn



韩先培(1984—),博士,研究员,主要研究领域为信息抽取、知识图谱、问答、阅读理解、自然语言处理等。
E-mail: xianpei@nfs.iscas.ac.cn



孙乐(1971—),博士,研究员,主要研究领域为信息抽取、知识图谱、问答、阅读理解、自然语言处理等。
E-mail: sunle@nfs.iscas.ac.cn