

# 第十一章 基于知识的问答

## 1. 任务定义、目标和研究意义

问答系统 (Question Answering, QA) 是指让计算机自动回答用户所提出的问题, 是信息服务的一种高级形式。不同于现有的搜索引擎, 问答系统返回用户的不再是基于关键词匹配的相关文档排序, 而是精准的自然语言形式的答案。华盛顿大学图灵中心主任 Etzioni 教授 2011 年曾在 Nature 上发表文章《Search Needs a Shake-Up》, 其中明确指出: “以直接而准确的方式回答用户自然语言提问的自动问答系统将构成下一代搜索引擎的基本形态”[Etzioni O., 2011]。因此, 问答系统被看做是未来信息服务的颠覆性技术之一, 被认为是机器具备语言理解能力的主要验证手段之一。因此, 对其开展研究具有重要的学术和实际意义。特别是近些年, 随着人工智能热潮到来, 无论是学术界还是产业界, 都给予其极大关注和投入。

纵观问答系统的技术演进, 其一直伴随的人工智能技术的发展而发展。近些年, 问答系统更是取得一系列倍受关注的成果。2011 年, IBM Watson 自动问答机器人在美国智力竞赛节目 Jeopardy 中战胜人类选手, 在业内引起了巨大的轰动。随着人工智能技术的突飞猛进, 各大 IT 巨头更是相继推出以问答系统为核心技术的产品和服务, 如移动生活助手 (Siri、Google Now、Cortana、小冰等)、智能音箱 (HomePod、Alexa、叮咚音箱等、公子小白等) 等, 这似乎让人们看到了黎明前的阳光, 甚至认为现有的问答技术已经十分成熟。

尽管 IBM Watson 系统在 Jeopardy 中战胜了人类选手, 但是其核心技术并没有突破传统基于“检索+抽取”的问答模式, 缺乏对于文本语义深层次的分析 and 处理, 难以实现知识的深层逻辑推理, 无法达到人工智能的高级目标。Watson 的成功也已经被证明仅仅局限于限定领域、特定类型的问题, 离语义的深度理解以及智能问答还有很大的距离, 其他问答系统, 如 Siri 等, 也存在同样的问题。因此, 面对已有问答模式的不足, 为了提升信息服务的准确性与智能性, 研究者近些年逐步把目光投向知识图谱 (Knowledge Graph)。其意图是通过信息抽取、关联、融合等手段, 将互联网文本转化为结构化的知识, 利用实体以及实体间语义关系对于整个互联网文本内容进行描述和表示, 从数据源头对于信息进行深度的挖掘和理解。同时, 互联网中已经有一些可以获取的大规模知识图谱, 例如 DBpedia[Lehmann et al., 2014]、Freebase[Bollacker, 2008]、YAGO[Suchanek et al., 2007]等。这些知识图谱多是以实体、关系为基本单元所组成的图结构。

基于这样的结构化的知识, 分析用户自然语言问题的语义, 进而在已构建的结构化知识图谱中通过检索、匹配或推理等手段, 获取正确答案, 这一任务称之

为知识库问答（Question Answering over Knowledge Base, KBQA）。这一问答范式由于已经在数据层面通过知识图谱的构建对于文本内容进行了深度挖掘与理解，能够有效地提升问答的准确性。

## 2. 研究内容和关键科学问题

知识库问答系统在回答用户问题时，需要正确理解用户所提的自然语言问题，抽取其中的关键语义信息，然后在已有单个或多个知识库中通过检索、推理等手段获取答案并返回给用户。其中所涉及的关键技术包括：词法分析、句法分析、语义分析、信息检索、逻辑推理、语言生成等。传统知识库问答系统多集中在限定领域，针对有限类型的问题进行回答。然而伴随大数据的飞速发展，已有知识图谱的规模在不断增大，所涉及的领域不断增多。现有研究趋向于开放域、面向大规模、开放域、多源异构知识库问答系统构建。总体来讲，主要面临如下三个关键科学问题。

### 2.1 问句语义解析

知识库问答要回答用户的问题，首先就要正确理解用户所提问题的语义内容。面对结构化知识库，需要将用户问题转化为结构化的查询语句，进而在知识图谱进行查询、推理等操作，获取正确答案。因此，对于用户问题的语义解析是知识库问答研究所面临的首要科学问题。具体过程需要分析用户问题中的语义单元与知识图谱中的实体、概念进行链接，并分析问句中这些语义单元之间的语义关系，将用户问题解析成为知识图谱中所定义的实体、概念、关系所组成的结构化语义表示形式。其中涉及词法分析、句法分析、语义分析等多项关键技术，需要自底向上从文本的多个维度理解其中包含的语义内容。在词语层面，需要在开放域环境下，研究实体(Entity)和术语(Terminology)的识别、答案类型词(Lexical Answer Type)识别、实体消歧(Entity Disambiguation)等关键技术。在句法层面，需要解析句子中词与词之间、短语与短语之间的句法关系，分析出句子句法结构。在语义层面，需要根据词语层面、句法层面的分析结果，将自然语言问句解析成可计算的结构化的逻辑表达形式（如一阶谓词逻辑表达式）。传统知识库问答方法面对单一领域有限规模知识图谱，多涉及的实体、概念、关系规模较小，通常采用模板、或者小规模机器学习算法进行语义解析。但是当面对大规模、多领域知识库时，随着实体、概念、关系规模增大，语义解析算法的复杂度也指数增加，如何获取实体提及，如何进行开放域关系抽取等问题仍然是学术界需要面对的难点问题。目前，已有一些工作利用深度神经网络将用户问题解析成为隐式表达的分布式数值向量的形式，其中蕴含的用户问句的关键语义，但是如何在分布式表示过程中与知识图谱相关联，反映其中所蕴含的实体、关系等关键语义也

是另一个科学问题。

## 2.2 大规模知识推理

在问答过程中，并不是所有的答案都能通过在知识图谱中进行检索或查询就可以获取答案。主要原因是已有知识库本身的覆盖度有限。需要在已有的知识体系中，通过知识推理的手段获取这些隐含的答案。例如，知识库中包括了一个人的“出生地”信息，但是没包括这个人的“国籍”信息，虽然知识库中对于人物对应了“国籍”属性，但是由于没有直接给出该属性的值，因此还是不能回答诸如“某某人是哪国人？”这样的问题。但是实际上我们都知道，一般情况下，一个人的“出生地”所属的国家就是他（她）的“国籍”。这些隐含知识天然存在于人的常识知识体系中，但在已有知识库中，并未被编码进去。面对知识库问答，就需要通过推理的方式学习到这样的模式。传统推理方法基于符号的知识表示形式，通过人工构建的推理规则推理出答案。但是面对大规模、开放域的问答场景，如何进行规则学习，如何解决规则冲突仍然是亟待解决的难点问题。目前，伴随深度学习的飞速发展，基于分布式表示的知识表示学习方法能够讲实体、概念以及它们之间的语义关系表示为低维空间中的对象（向量、矩阵等），通过在低维空间中的数值计算完成知识推理任务。虽然就目前来说，这类推理的效果离实用还有段距离，但是我们认为这是值得探寻的方法，特别是如何将已有的基于符号表示的逻辑推理与基于分布式表示的数值推理相结合，研究融合符号逻辑和表示学习的知识推理技术，是知识推理任务中的关键难点问题。

## 2.3 异构知识关联

由于用户问题的复杂性和多样性，问题的答案往往不能够在单一知识库中找到，需要综合多个知识库（多种语言、多种领域、多种模态）内的知识才能给出答案[Bizer et al., 2009]。例如

“谁出演了变形金刚并且和《Monkey Business》的演唱者结婚了？”

“谁出演了变形金刚”的信息需要在电影知识库中搜寻答案；而有关“结婚”的信息通常位于人物知识库中；“《Monkey Business》的演唱者”信息则位于音乐知识库中。因此，回答这个问句，需要综合电影、人物以及音乐三个不同知识库的信息，才能推出最终的答案：“乔什·杜哈明”。由于多源知识库之间存在结构差异、内容差异、语言差异、模态差异，要完成这一任务并不简单。（1）在面向多源异构知识库问答过程中，相对于面向单一知识库的问答，问句文本歧义更加严重。同一短语，在不同知识库中会映射为更多的概念（实体、关系）候选，这使得问句的语义解析更加困难。（2）问句中不同的子问题需要在不同的知识库中进行求解，这需要问答系统对于子问题进行精准的划分，同时确定子问题求解范围。（3）

不同源异构知识库之间存在冗余关联，不同知识库中的不同实体、关系间具有同指关系。多知识库问答需要利用这种同指关系对于多个知识库中的知识进行综合，从而回答用户问题。然而多源异构知识库间的同指关系通常并没有显式给出，是一种隐含关系。因此，系统需要挖掘知识库间的同指关系，完成异构知识库的关联与对齐，这对于构建多源异构知识库的问答系统有着重要的作用。

### 3. 技术方法和研究现状

根据技术路线的不同，已有知识库问答技术大致可以分为两类：1) 基于语义解析 (Semantic Parsing) 的知识库问答方法；2) 基于检索排序的知识库问答方法。下面将分别简要介绍技术现状。

#### 3.1 基于语义解析的知识库问答方法

在结构化数据形式的知识图谱上进行查询、匹配、推理等操作，最有效的方式是利用结构化的查询语句，例如：SQL、SPARQL 语句等。然而，这些语句通常是由专家编写，对于普通用户来说，自然语言仍然是最直接的交互方式。因此，如何把用户的自然语言问句转化为结构化的知识库查询语句便是进行问答的核心所在，其关键是对自然语言问句进行解析（如图 1 所示）。目前，主流方法是通过语义解析，将用户的自然语言问句转化成结构化的语义表示，例如 $\lambda$ 范式 [Kwiatkowski, et al., 2011] 和 DCS-Tree [Liang, et al., 2011]。相对应的也提出了很多语义解析语法或方法，例如组合范畴语法 (Category Compositional Grammar, CCG) [Kwiatkowski, et al., 2011] 以及依存组合语法 (Dependency-based Compositional Semantics, DCS) [Liang, et al., 2011] 等。

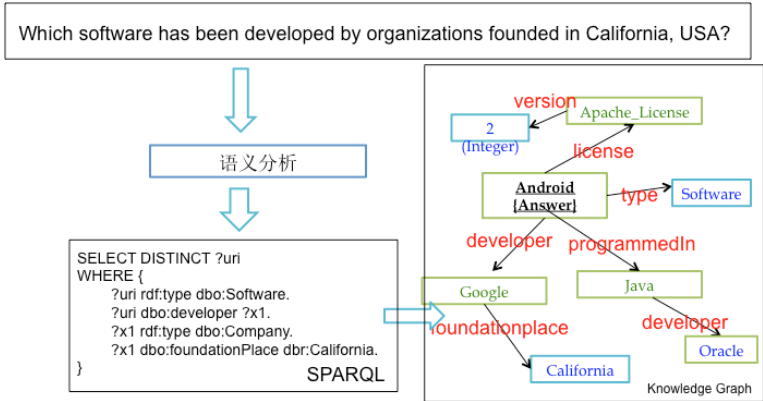


图 1. 基于语义解析的知识库问答过程

尽管很多语义解析方法在限定领域内能达到很好的效果，在这些工作中，很多重要组成部分（比如 CCG 中的词汇表和规则集）都是人工编写的。但是，当面对大规模多源异构知识库，传统的语义分析方法存在以下几个缺陷：（1）资源（例如词汇表、规则集）标注费时费力，传统方法在有限的训练数据下性能大打

折扣；(2) 大规模知识库的开放域特性使得文本歧义问题更加严重，传统语义分析方法难以处理这一问题；(3) 在很多场景下，回答一个问题需要多个知识库的综合运用。然而，不同来源知识库是异构的，它们在结构和内容都具有异质性 [Lopez, et al., 2011]，如何处理多知识库间的冗余和差异性，是面向多知识库的问答系统面临的主要挑战。因此，很多研究者针对上述问题进行研究，取得一系列进展。在面对训练数据标注困难这一挑战时，目前主要采用半监督学习 [Clarke et al., 2010]、外部资源 [Cai, et al., 2013]、弱监督学习 [Liang, et al., 2011] [Berant, et al., 2013]、同义对应 [Fader, et al., 2013] 等手段。在面对歧义更加严重的挑战时，主要采用联合模型 [Lu, 2014]、图搜索 [Tran, et al., 2009] 等手段。在面对多源异构知识库时，主要采用子问题划分 [Fader, et al., 2014] [Lopez, et al., 2012]、整数线性规划 [Zhang, et al., 2016] 等手段。

### 3.2 基于检索排序的知识库问答方法

但是，基于语义解析的知识库问答系统的处理范式通常仍然是基于符号逻辑的，缺乏灵活性。在分析问句语义过程中，易受到符号间语义鸿沟影响。同时从自然语言问句到结构化语义表达需要多步操作，多步间的误差传递对于问答的准确度也有很大的影响。近年来，深度学习技术以及相关研究飞速发展，越来越多的研究者开始研究深度学习技术在自然语言处理问题中的应用，例如情感分析、机器翻译、句法分析等等。知识库问答系统也不例外，已有相关的研究工作包括 [Bordes, et al., 2014; Bordes, et al., 2014; Dong, et al., 2015; Hao, et al., 2017]。与传统基于符号的知识库问答方法相比，基于表示学习的知识库问答方法更具鲁棒性，其在效果上已经逐步超过传统方法，如图 2 所示。这些方法的基本假设是把知识库问答看做是一个语义匹配的过程。通过表示学习，我们能够用户的自然语言问题转换为一个低维空间中的数值向量(分布式语义表示)，同时知识库中的实体、概念、类别以及关系也能够表示成为同一语义空间的数值向量。那么传统知识库问答任务就可以看成问句语义向量与知识库中实体、边的语义向量相似度计算的过程。

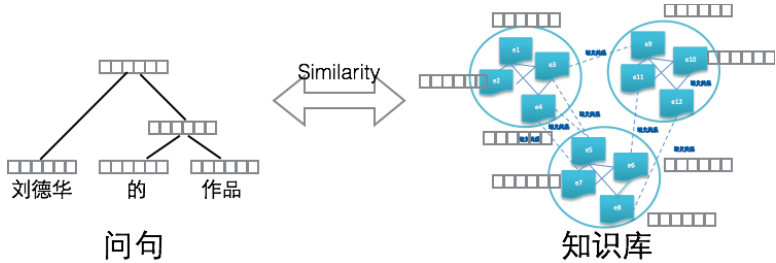


图 2、基于表示学习的知识库问答示意图

### 3.3 技术现状

已有的评测主要针对于一些限定领域的知识库进行问答。已有方法也取得了不错的结果。比如,在 Geoquery<sup>28</sup>(美国地理知识查询)数据集上(600 个训练样本, 280 个测试样本)上,使用 CCG 和本体匹配的方法 F 值能达到 89.0%,使用 DCS 的方法 F 值能达到 91.1%;在求职(JOBS)数据集上(500 个训练样本, 140 个测试样本),使用 CCG 的方法 F 值能达到 79.3%,使用 DCS 的方法 F 值能达到 95%。在这一方面, QALD (Question Answering over Linked Data) 评测的举办更是推动了这方面的研究。QALD 每年举办一届,目前已经举办到了第六届。每一次评测,组织者都会给出一些问题,要求参加评测系统在给定知识库的基础上,将所给问题转化为结构化的 SPARQL 查询语句,并在给定知识库上查询答案。但是,目前的研究趋势是从限定领域的知识库向大规模开放域甚至是多领域知识库进行扩展,例如 Freebase。与限定领域知识库相比,大规模开放知识库包含的资源 and 关系数量要大得多,比如 Geoquery 中只包含 8 个关系谓词,而 Freebase 包含上万个关系。因此开放知识库上的语义解析效果有明显的下降。例如利用 Freebase 知识库,开放查询测试的最好的效果只有 39.9% [Berant, 2014];而在 QALD 评测中,在 DBpedia 上、开放查询中,表现最好的问答系统的正确率只有 40% [He, et al., 2014]。下图给出在面对开放域知识库 Freebase 时,在公开问题库 WebQuestion 上,已有系统能够达到的精度。

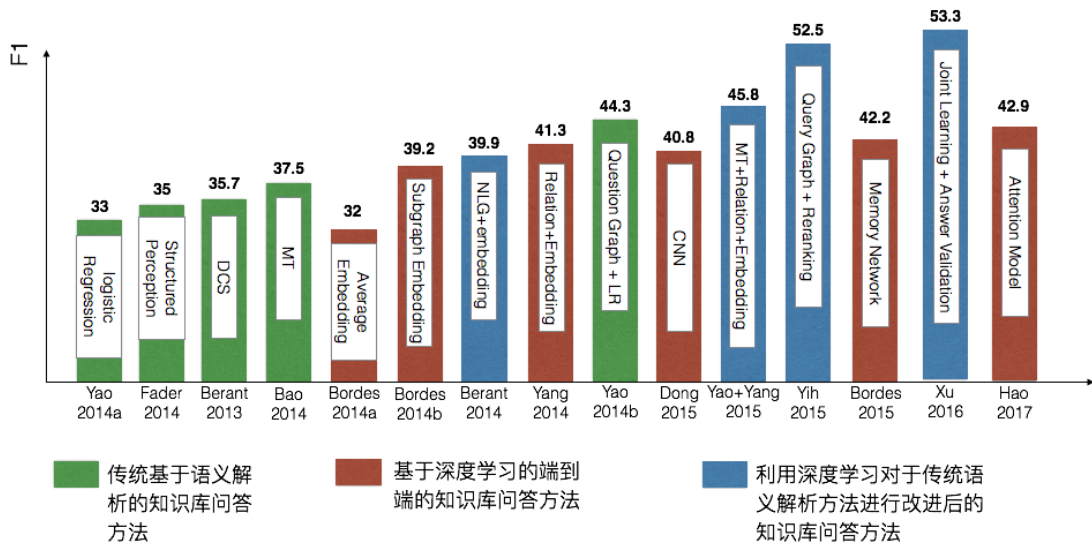


图 3. 已有知识库问答方法在 WebQuestion 问题集上的性能

## 4. 技术展望与发展趋势

纵观知识库问答研究发展的态势和技术现状,以下研究方向或问题将可能成

<sup>28</sup> <http://www.cs.utexas.edu/users/ml/nldata/geoquery.html>

为未来整个领域和行业重点关注的方向：

- 面向复杂问句的深度学习知识库问答方法

在实际问答场景下，用户的问题往往是复杂问句，其中所包含的语义关系复杂多样，子问题嵌套、文本歧义现象尤为突出。然而，已有基于深度学习的知识库问答方法目前尚只能解决简单类型问题（包含单一关系的问题类型）。在面对复杂问题时，例如有限制条件的问题（what did obama do before he was elected president?）、聚合问题（when's the last time the steelers won the superbowl?）等，已有方法处理手段单一，常忽略知识库与文本语义的关联与约束，缺乏在已有知识资源约束下的文本语义表示学习手段。因此，如何利用深度学习的方法解决复杂问题值得继续关注。

- 分布式表示与符号表示相结合的知识库问答

目前，基于深度学习的知识库问答方法试图通过高质量已标注的问题-答案建立联合学习模型，同时学习知识库库和问题的语义表示及他们之间的语义映射关系，试图通过分布式表示（向量）间的简单数值运算对于复杂的问答过程进行建模。这类方法的优势在于把传统的问答语义解析的复杂步骤转变为一个可学习的过程，虽然取得了一定的效果，但是训练过程容易受到训练数据质量的影响，缺乏已有知识的约束。同时，问答过程也缺乏可解释性。从目前自然语言处理很多任务来看，将统计与知识相结合是未来技术的重点突破方向。知识库问答也不例外，目前已经有将基于分布式表示的方法（深度学习）与传统基于符号表示的语义解析方法相结合的初步工作，例如神经图灵机[Liang, et al., 2017]。但是工作还很初步。如何将深度学习与传统语义方法进行深度融合，使这两种技术路线相互融合、相互约束，提升知识库问答的效果，是一个很值得深入研究的方向。

- 面向问答的深度推理

尽管已有知识图谱规模已经十分巨大，能够覆盖多个领域，但仍旧面临信息缺失的现象，这对于知识库问答带来巨大的挑战。这就需要面向问答的深度推理。传统基于符号逻辑的逻辑推理方法基于严格的符号匹配，过分依赖于推理规则的生成，因此具有领域适应性差、无法进行大规模推理的缺点。而深度学习基于分布式语义表示，利用语义空间中的数值模糊计算替代传统问答过程中的符号严格匹配，为解决上述问题提供一种途径，但也存在推理结果准确度低、可解释性差的问题。因此，如果利用深度学习大规模、可学习的特点，在深度神经网络框架下，融入传统的逻辑推理规则，构建精准的大规模知识推理引擎是自动问答迫切需要解决的难点问题。

- 对话中的自然语言形式回复

传统的自动问答都是采用一问一答的形式。然而在很多场景下，需要提问者和系统进行多轮对话交互，实现问答过程。这时，需要系统返回用户的答案不再



只是单一实体、概念、关系的形式，而是需要是以自然语言的形式返回答案。这就需要自动生成自然语言的回复。现有方法多利用 `sequence-to-sequence` 模型进行自然语言生成，在这一过程中，如何与知识库相结合，将知识库问答的答案加入自然语言回复中，仍是亟待解决的问题。

总之，自动问答作为人工智能技术的有效评价手段，已经研究 60 余年了。整体上，知识库问答技术的发展趋势是从限定领域向开放领域、从单个数据源向多个数据源、从浅层语义分析向深度推理不断推进。我们有理由相信，随着自然语言处理、深度学习、知识工程等相关技术的飞速发展，知识库问答技术在未来有可能得到相当程度的突破。伴随着更多 AI 应用的实际落地，我们期待看到这一技术将在不远的未来得到更大、更广的应用。

## 参考文献

- [Bizer C., et al., 2009] C. Bizer, T. Heath, and T. Berners-Lee. Linked data - the story so far. *Int. J. Semantic Web Inf. Syst.*, 5(3):1--22, 2009.
- [Berant J, et al., 2014] Berant J, Liang P. Semantic parsing via paraphrasing[C]. In *Proceedings of ACL*. 2014, 7(1): 92.
- [Bollacker, 2008] K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor, Freebase: a collaboratively created graph database for structuring human knowledge, in *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, 2008, pp. 1247–1250.
- [Bordes, et al., 2014] Bordes A, Weston J, Usunier N. Open question answering with weakly supervised embedding models, *Machine Learning and Knowledge Discovery in Databases*. Springer Berlin Heidelberg, 2014: 165-180.
- [Bordes, et al., 2014] Bordes A, Chopra S, Weston J. Question Answering with Subgraph Embeddings In *Proceedings of EMNLP 2014*.
- [Cai Q, et al., 2013] Cai Q, Yates A. Large-scale Semantic Parsing via Schema Matching and Lexicon Extension[C]. *ACL (1)*. 2013: 423-433.
- [Dong, et al., 2015] Dong L, Wei F, Zhou M, et al. Question Answering over Freebase with Multi-Column Convolutional Neural Networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*. 2015, 1: 260-269
- [Etzioni O., 2011] O. Etzioni, Search needs a shake-up, *Nature*, vol. 476, no. 7358, pp. 25–26, 2011.
- [Fader A, et al., 2013] Fader A, Zettlemoyer L S, Etzioni O. Paraphrase-Driven Learning for Open Question Answering. In *Proceedings of ACL 2013*: 1608-1618.
- [Fader, et al., 2014] Bordes A, Chopra S, Weston J. Question Answering with Subgraph Embeddings In *Proceedings of EMNLP 2014*
- [Hao, et al., 2014] Yanchao Hao, Yuanzhe Zhang, Shizhu He, Kang Liu and Jun Zhao, An End-to-End Model for Question Answering over Knowledge Base with Cross-



- Attention Combining Global Knowledge, in Proceedings of ACL 2017, Vancouver, Canada, July 30-August 4.
- [He, 2014] He S, Liu K, Zhang Y, Xu LH, Zhao J. Question Answering over Linked Data Using First-order Logic. In Proceedings of EMNLP. 2014
- [Kwiatkowski, et al., 2011] T. Kwiatkowski, L. Zettlemoyer, S. Goldwater, and M. Steedman, Lexical generalization in ccg grammar induction for semantic parsing, in Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, 2011, pp. 1512–1523. <sup>[1]</sup><sub>[SEP]</sub>
- [Lehmann et al., 2014] J. Lehmann, R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P. N. Mendes, S. Hellmann, M. Morsey, P. van Kleef, S. Auer et al., Dbpedia a large-scale, multilingual knowledge base extracted from wikipedia, Semantic Web, 2014.
- [Liang, et al., 2011] P. Liang, M. I. Jordan, and D. Klein, Learning dependency- based compositional semantics, in Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies- Volume 1, 2011, pp. 590–599.
- [Liang, et al., 2017] Chen Liang, Jonathan Berant, Quoc Le, Kenneth D. Forbus, Ni Lao, Neural Symbolic Machines: Learning Semantic Parsers on Freebase with Weak Supervision , In Proceedings of ACL 2017
- [Lopez V, et al., 2012] Lopez V, Fernández M, Motta E, et al. Poweraqua: Supporting users in querying and exploring the semantic web[J]. Semantic Web, 2012, 3(3): 249-265.
- [Lu, 2014] Lu W. Semantic Parsing with Relaxed Hybrid Trees[C]. In Proceedings of EMNLP. 2014.
- [Suchanek et al., 2007] F. M. Suchanek, G. Kasneci, and G. Weikum, Yago: a core of semantic knowledge, in Proceedings of the 16th international conference on World Wide Web, 2007. <sup>[1]</sup><sub>[SEP]</sub>
- [Tran T, et al., 2009] Tran T, Wang H, Rudolph S, et al. Top-k exploration of query candidates for efficient keyword search on graph-shaped (rdf) data[C]. Data Engineering, 2009. ICDE'09. IEEE 25th International Conference on. IEEE, 2009: 405-416.
- [Zhang, et al., 2016] Yuanzhe Zhang, Shizhu He, Kang Liu and Jun Zhao, A Joint Model for Question Answering over Multiple Knowledge Bases, in Proceedings of AAAI 2016, Phoenix, USA, February, 12-17.