赛尔原创 | 知识图谱的发展概述

原创 2017-10-16 姜天文 哈工大SCIR

作者:哈工大SCIR博士生姜天文

“知识图谱（Knowledge Graph）”的概念由Google公司在2012年提出[1]，是指其用于提升搜索引擎性能的知识库。与近年来其他学者相同，本文中的“知识图谱”泛指知识库项目，而非特指Google的知识图谱项目。

知识图谱的出现是人工智能对知识需求所导致的必然结果，但其发展又得益于很多其他的研究领域，涉及专家系统、语言学、语义网、数据库，以及信息抽取等众多领域，是交叉融合的产物而非一脉相承。

图1 多领域共同促进知识图谱发展

知识图谱的早期发展

早在上个世纪70年代，专家系统（**Expert Systems**）作为人工智能的重要分支，是指利用知识和推理过程来解决那些借助人类专家知识才能得已解决的问题的计算机程序[2]。八十年代，专家系统的发展激增，日本的五代机项目就是在这期间开始的，专家系统是其核心部分。专家系统一般由两部分组成：知识库与推理引擎。人类专家提供知识，再将这种显式的知识映射并存储到知识库中用来推理。

Cyc是这一期间较为出色的项目[3]，由Douglas Lenat在1984年设立，旨在收集生活中常识知识并将其编码集成到一个全面的本体知识库。Cyc知识库中的知识使用专门设计的**CycL**进行表示。同其他专家系统一样，Cyc不仅包括知识，而且提供了非常多的推理引擎，支持演绎推理和归纳推理。目前Cyc知识库涉及50万条概念的500万条常识知识。**OpenCyc**是其开放出来免费供大众使用的部分知识，包括24万条概念的约240万条常识知识。

对词汇的理解是解读自然语言的关键，语言学家所创造的词典为人类而非机器的阅读提供了便利，虽然有电子词典的存在，但机器仍无法很好的从中获取词汇含义。1985年，普林斯顿大学认识科学实验室在心理学教授乔治·A·米勒的指导下开始建立和维护名为**WordNet**的英语字典[4]，旨在为词典信息和现代计算提供更加有效的结合，为计算机程序提供可读性较强的在线词汇数据库。在WordNet中，名词、动词、形容词以及副词被按照认知上的同义词分组，称为**synsets**，每一个synset表征一个确定的概念。synset之间通过概念语义以及词汇关系链接。在汉语中，类似的典型代表有《同义词词林》[15]及其扩展版[16]、知网（**HowNet**）[17]等，都是从语言学的角度，以概念为最基本的语义单元构建起来的可以被计算机处理的汉语词典。

这些早期的知识图谱都是利用相关领域专家进行人工构建，具有很高的准确率和利用价值，但是其构建过程耗时耗力而且存在覆盖性较低的问题。

链接数据与基于百科知识的知识图谱构建

1989年万维网的出现，为知识的获取提供了极大的方便，1998年，万维网之父蒂姆·伯纳斯·李再次提出语义网（**Semantic Web**），其初衷是让机器也同人类一样可以很好地获取并使用知识[5,6,7]。不同于人工智能中训练机器使之拥有和人类一样的认知能力，语义网直接向机器提供可直接用于程序处理的知识表示[5]。但语义网是一个较为宏观的设想并且其设计模型是“自顶向下”的，导致其很难落地，学者们逐渐将焦点转向数据本身。2006年，伯纳斯·李提出链接数据（**Linked Data**）的概念，鼓励大家将数据公开并遵循一定的原则（2006年提出4条原则，2009年精简为3条原则）将其发布在互联网中[8,9]，链接数据的宗旨是希望数据不仅仅发布于语义网中，而需要建立起数据之间的链接从而形成一张巨大的链接数据网。其中，最具代表性的当属2007年开始运行的**DBpedia**项目[10]，是目前已知的第一个大规模开放域链接数据。

DBpdia项目最初是由柏林自由大学和莱比锡大学的学者发起的，其初衷是缓解语义网当时面临的窘境，第一份公开数据集在2007年时发布，通过自由授权的方式允许他人使用。Leipzig等学者[10]认为在大规模网络信息的环境下传统“自上而下”地在数据之前设计本体是不切实际的，数据及其元数据应当随着信息的增加而不断完善。数据的增加和完善可以通过社区成员合作的方式进行，但这种方式涉及数据的一致性、不确定性，以及隐式知识的统一表示等诸多问题。Leipzig等人[10]认为探寻这些问题最首要并高效的方式就是提供一个内容丰富的多元数据语料，有了这样的语料便可以极大推动诸如知识推理、数据的不确定管理技术，以及开发面向语义网的运营系统。朝着链接数据的构想，DBpedia知识库利用语义网技术，如资源描述框架（RDF）[18]，与众多知识库（如WordNet、Cyc等）建立链接关系，构建了一个规模巨大的链接数据网络。

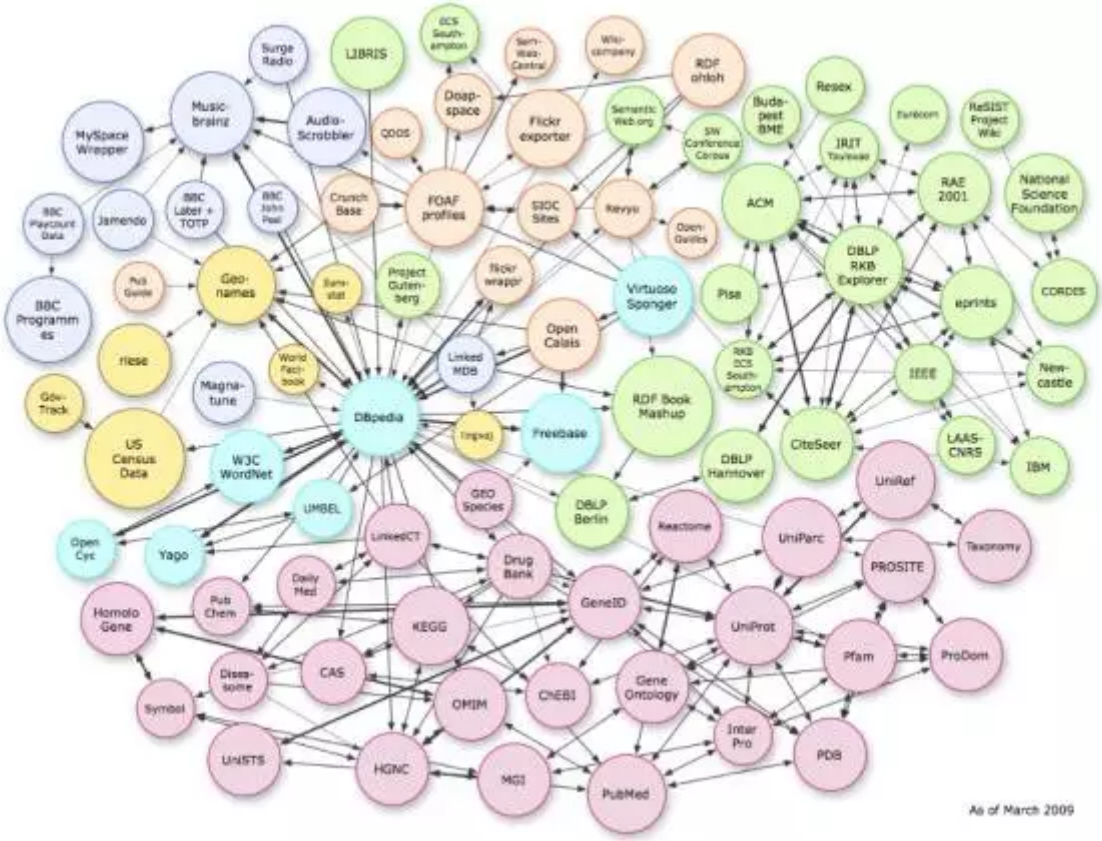

图2 以DBpedia为核心的链接数据网络

2001年，一个名为维基百科（Wikipedia）的全球性多语言百科全书协作计划开启[11]，其宗旨是为全人类提供自由的百科全书，在短短几年的时间里利用全球用户的协作完成数十万词条（至今拥有上百万词条）知识。维基百科的出现推动了很多基于维基百科的结构化知识的知识库的构建，**DBpedia**[10]、**Yago**[12]等都属于这一类知识库。

Yago是由德国马普研究所于2007年开始的项目，针对当时的应用仅使用单一源背景知识的情况，建立了一个高质量、高覆盖的多源背景知识的知识库。前面介绍的专家构建的WordNet拥有极高的准确率的本体知识，但知识覆盖度仅限于一些常见的概念或实体；相比之下，维基百科蕴含丰富的实体知识，但维基百科多提供的概念的层次结构类似标签结构并不精确，直接用于本体构建并不适合。Yago的主要思路是将WordNet与维基百科二者的知识结合，即利用WordNet的本体知识补充维基百科中实体的上位词知识，从而获取大规模高质量、高覆盖的知识库。截至目前，Yago拥有超过1千万实体的1.2亿条事实知识，同时近些年也构建起了与其他知识库的链接关系。

DBpedia主要通过社区成员来定义和撰写准确的抽取模版，从维基百科中抽取结构化信息（如，infobox）构建大规模知识库，另外本体（即知识库的元数据、schema）的构建也是通过社区成员合作完成的。由于维基百科是社区撰写，其知识表达难免有不一致的情况，DBpedia利用mapping技术与抽取模版来实现知识描述的统一与一致性。另外，为了实现知识的更新与扩增，DBpedia开发DBpediaLive来保持与维基百科的同步。在2016年发行的版本中，DBpedia拥有超过6百万实体及其数十亿事实知识，其中人工构建的本体库包含760种类别信息。同时，DBpedia拥有大量的跨语言知识，共拥有除英语外的66亿其他语言事实知识。

2007年，**Freebase**[13]开始构建，类似维基百科，其内容主要来自其社区成员的贡献，但与维基百科最大的不同之处在于Freebase中都是结构化的知识，在维基百科中人们编辑的是文章，而在Freebase中编辑的是知识。在Freebase中，用户是其主要核心，除了对实体的编辑，用户也参与本体库的构建、知识的校对，以及与其他知识库的链接工作。除人工输入知识，Freebase也主动导入知识，如维基百科的结构化知识。Freebase拥有大约2千万实体，目前被Google公司收购，Freebase的API服务已经关闭但仍提供数据的下载。

2012年，考虑到维基百科中大部分的知识都是非结构组织起来的，带来诸多问题（如：无法对知识进行有效的搜索与分析，进而知识无法得到很好的重用，甚至存在知识的不一致性的现象），维基媒体基金会推出**Wikidata**项目[14]，一个类似于Freebase的大规模社区成员合作知识哭，旨在用一种全新的方式管理知识以克服以上的存在于维基百科中的问题。

以上所介绍的知识图谱都是基于英文语言的，即使是多语言知识图谱也是以英文为主语言，其他语言知识是用过跨语言知识（如，语言间链接（ILLs）、三元组对齐（TWA））链接得到。近些年，国内推出了大量以中文为主语言的知识图谱，它们主要都是基于百度百科和维基百科的结构化信息构建起来的。如上海交通大学的**zhishi.me**[19]、清华大学的**XLore**[20]、复旦大学的**CN-pedia**[21]。2017年，由国内多所高校发起**cnSchema.org**项目[23]，旨在利用社区力量维护开放域知识图谱的Schema标准。

基于自由文本的开放域知识图谱构建

上述介绍的知识图谱的构建方式包括人工编辑和自动抽取，但自动抽取方法主要是基于在线百科中结构化信息而忽略了非结构化文本，而互联网中大部分的信息恰恰是以非结构化的自由文本形式呈现。与链接数据发展的同期，很多基于信息抽取技术的知识获取方法被提出，用以构建基于自由文本的开放域知识图谱。

2007年，华盛顿大学Banko等人[24]率先提出开放域信息抽取（**OIE**），直接从大规模自由文本中直接抽取实体关系三元组，即头实体、关系指示词，以及尾实体三部分，类似于语义网中RDF规范的SPO结构。在OIE提出之前，也有很多面向自由文本的信息抽取被提出，但这些方法主要的思路都是为每个目标关系训练相应的抽取器。这类传统的信息抽取方法在面对互联网文本中海量的关系类别时无法高效工作，即为每个目标关系训练抽取器时不现实的，更为严重的是很多情况下面对海量的网络文本我们无法事先明确关系的类型。OIE通过直接识别关系词组（**relation phrases**）也称关系指示词，即显式表证实体关系的词组，来抽取实体关系。基于OIE的指导思想，华盛顿大学陆续推出**TextRunner**[24]、**Reverb**[25]、**OLLIE**[26]等基于自由文本的开放域三元组抽取系统；以及卡耐基梅隆大学的**NELL**系统[27,29]、德国马普研究中心的**PATTY**等[28]。这些系统有的需要自动构造标注的训练语料，进而从中提取关系模版或训练分类器；有的则依据语法或句法特征直接从分析结果中抽取关系三元组。接下来，本文将简要介绍下具有代表性的Reverb和NELL系统的实现思想。

Reverb针对之前的OIE系统中存在的两个问题：不连贯抽取与信息缺失抽取，提出句法约束：对于多词语关系词组，必须以动词开头、以介词结束，并且是由句子中毗邻的单词组成。该约束可以有效缓解以上两个问题造成的抽取失败。进一步，为了避免由句法约束带来的冗长

的并且过于明确的关系指示词，Reverb引入了启发式的词法约束。总的来说，Reverb提出了两个简单却高效的约束，在面向英文自由文本的开放域知识抽取中取得了不错的效果，很具启发意义。

never-ending learning 被定义为是一种不同于传统的机器学习方式[29]，通过不断地阅读获取知识，并不断提升学习知识的能力以及利用所学知识进行推理等逻辑思维。**NELL**就是一种这样的智能体，其任务是学习如何阅读网页以获取知识。

NELL的输入有：

- 定义了类别和二元关系的初始本体库；
- 对于每个类别和关系的训练种子数据；
- 网页数据（从预先准备好的网页集合中获取、每天从*Google*搜索*API*获取）；
- 偶尔的人工干预，

NELL每天24小时不停歇的进行如下操作：

- 从网页中阅读（抽取）知识事实用以填充知识库，并移除之前存在于知识库中不正确知识事实，每个知识具有一定的置信度以及参考来源；
- 学习如何比前一天更好地阅读（抽取）知识事实，

NELL从2010年1月开始进行上述阅读过程，目前所产生的知识库已经拥有超过8千万的相互链接的事实，以及上百万学习到的短语。

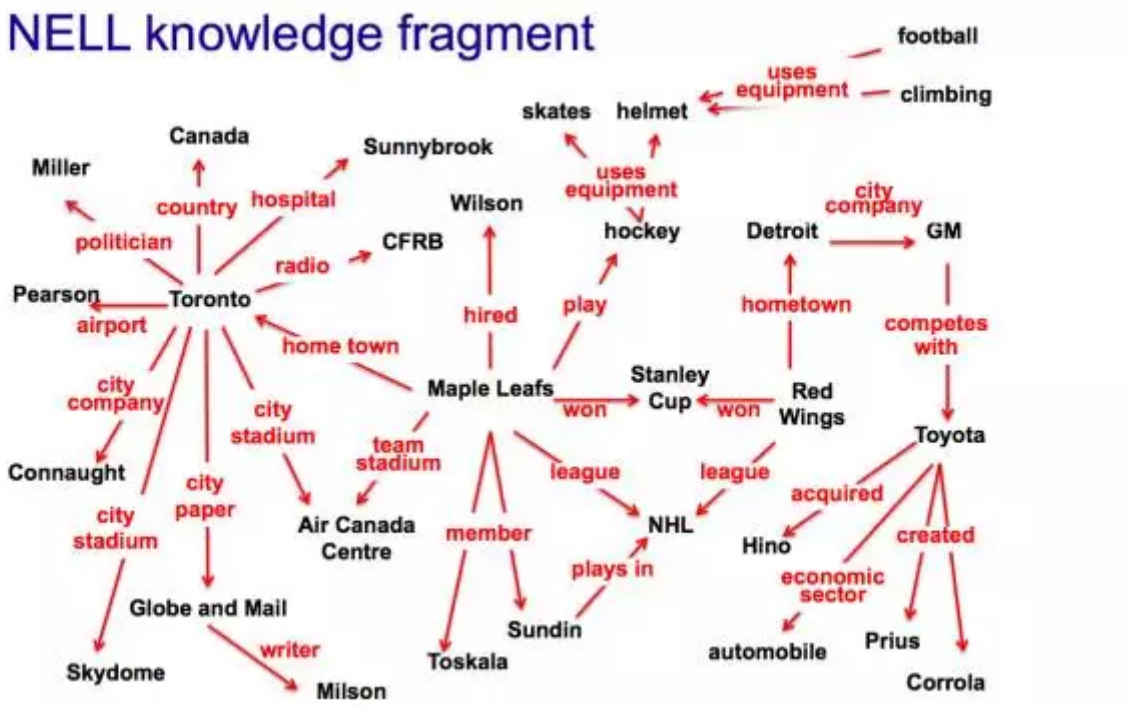

图3 NELL knowledge fragment

自动构建的开放域中文实体知识图谱——《大词林》

上述所介绍的OIE系统大多专注于对开放域实体关系三元组的抽取，但忽略了对于知识图谱不可或缺的同时也是至关重要的本体库的构建，即知识图谱元数据或称为Schema的构建，是为三元组赋以语义的关键。2014年，由哈尔滨工业大学社会计算与信息检索研究中心发起的《大词林》项目，面向包括自由文本的多信息源对实体的类别信息进行自动抽取并层次化，进而实现对实体上下位关系体系的自动构建，而上下位关系体系正是本体库的核心组成之一。

《大词林》的构建不需要领域专家的参与，而是基于多信息源自动获取实体类别并对可能的多个类别进行层次化，从而达到知识库自动构建的效果。同时也正是由于《大词林》具有自动构建能力，其数据规模可以随着互联网中实体词的更新而扩大，很好地解决了以往的人工构建知识库对开放域实体的覆盖程度极为有限的问题。

另外，相比以往的类别体系知识库，《大词林》中类别体系的结构也更加灵活。如《同义词词林（扩展版）》中每个实体具有具备五层结构，其中第四层仅有代码表示，其余四层由代码和词语表示，而《大词林》中类别体系结构的层数不固定，依据实体词的不同而动态变化，如“哈工大”一词有7层之多，而“中国”一词有4层；另外，《大词林》中的每一层都是用类别词或实体词表示。

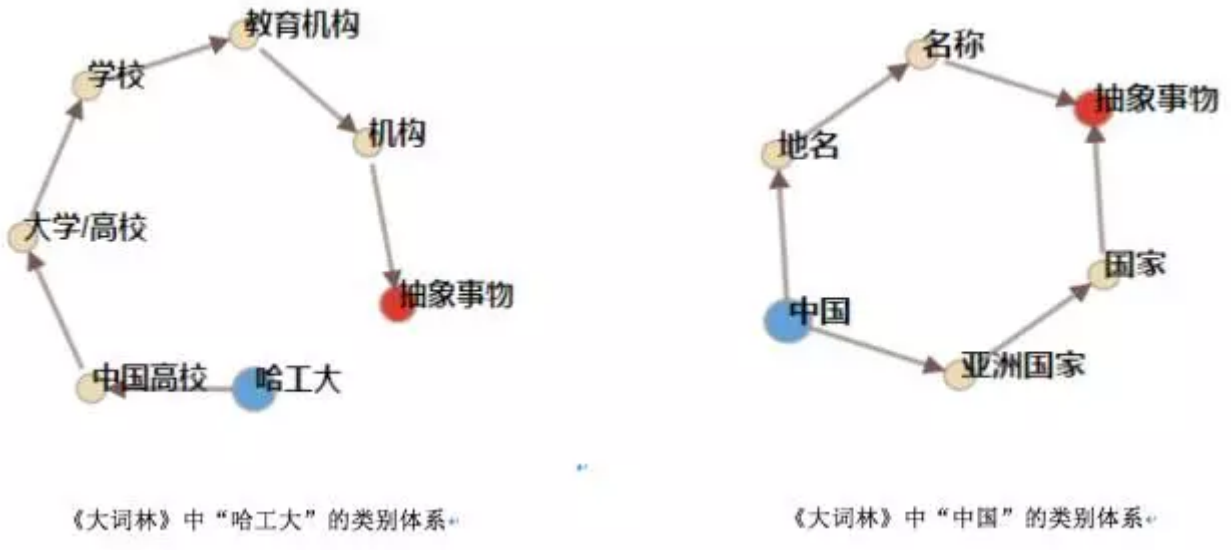

图4《大词林》中“哈工大”的类别体系图（左）和“中国”的类别体系图（右）

自2014年11月27日上线，《大词林》不断添加中文实体及其层次化类别信息，自动构建开放域实体知识库。目前，《大词林》中包括约900万实体、约17万类别；平均每个命名实体有1.77个不同粒度的优质类别；上下位关系超过1千万对，其中实体与上位词之间的上下位关系与上位词之间的上下位关系准确率均达到90%以上。

《大词林》（http://www.bigcilin.com/）系统网站支持用户查询任意实体，并以有向图的形式展现实体的层次化类别，同时支持以目录方式供用户浏览部分公开的知识库。人工智能中关键的一步是知识的获取与构建，《大词林》作为基于上下位关系的中文知识库，随着互联网中实体词的增加不断扩充其数据规模，并即将加入实体间关系、实体属性等网状关系结构，这对于基于知识库的智能系统无疑是一笔巨大的宝藏。

参考文献

[1] Singhal A. Introducing the knowledgegraph: things, not strings[J]. Official google blog, 2012.

[2] Feigenbaum E A. Expert systems in the1980s[J]. State of the art report on machine intelligence.Maidenhead:Pergamon-Infotech,1981.

[3] Lenat D B, Prakash M, Shepherd M. CYC:Using common sense knowledge to overcome brittleness and knowledge acquisition bottlenecks[J]. AI magazine, 1985, 6(4): 65.

[4] Miller G A. WordNet: a dictionary browser[J]. Information in Data, 1985: 25-28.

[5] Berners-Lee T. Semantic web roadmap[J]. 1998.

[6] Berners-Lee T, Hendler J, Lassila O.The semantic web[J]. Scientific american, 2001, 284(5): 28-37.

[7] Shadbolt N, Berners-Lee T, Hall W. Thesemantic web revisited[J]. IEEE intelligent systems, 2006, 21(3): 96-101.

[8]Berners-Lee T.Linked data-designissues[J]. http://www.w3.org/DesignIssues/LinkedData.html, 2006.

[9] Berners-Lee T. The next web[J]. TED.com, 2009.

[10] Auer S, Bizer C, Kobilarov G, et al.Dbpedia: A nucleus for a web of open data[J]. The semantic web, 2007: 722-735.

[11] Wales J, Sanger L. Wikipedia: The free encyclopedia[J]. Accessed via http://en.wikipedia.org/wiki/Main_Page (27 November 2011), 2001.

[12] Suchanek F M, Kasneci G, Weikum G.Yago: a core of semantic knowledge[C] //Proceedings of the 16th international conference on World Wide Web. ACM, 2007: 697-706.

[13] Bollacker K, Cook R, Tufts P.Freebase: A shared database of structured general human knowledge[C]//AAAI.2007, 7: 1962-1963.

[14] Vrandečić D. Wikidata: a new platformfor collaborative data collection[C] //Proceedings of the 21st International Conference on World Wide Web. ACM, 2012: 1063-1064.

[15] 梅家驹. 同义詞詞林[M]. 上海辞书出版社,1983.

[16] 《同义词词林（扩展版）》：https://www.ltp-cloud.com/download/

[17] 董振东, 董强. 知网简介[J].1999-09-23.[2004-03-06]. http://www.keenage.com, 1999.

[18] Swick R R. Resource Description Framework (RDF) Model and Syntax Specification W3C Recommendation[J]. W3c Recommendation World Wide Web Consortium, 1999.

[19] Niu, X.; Sun, X.; Wang, H.; Rong, S.;Qi, G.; and Yu, Y. 2011. Zhishi. me-weaving chinese linking open data. TheSemantic Web–ISWC 2011 205–220.

[20] Wang, Z.; Li, J.; Wang, Z.; Li, S.;Li, M.; Zhang, D.; Shi, Y.; Liu, Y.; Zhang, P.; and Tang, J. 2013. Xlore: Alarge- scale english-chinese bilingual knowledge graph. In Pro- ceedings of the 2013th International Conference on Posters & Demonstrations Track-Volume1035, 121–124. CEUR- WS. org.

[21] Xu, B.; Xu, Y.; Liang, J.; Xie, C.;Liang, B.; Cui, W.; and Xiao, Y. 2017. Cn-dbpedia: A never-ending chinese knowledge extraction system. In International Conference on In- dustrial,Engineering and Other Applications of Applied In- telligent Systems, 428–438.Springer.

[22] 《大词林》项目官网：http://www.bigcilin.com

[23] cnSchema官网：http://cnschema.org

[24] Michele Banko, Michael J Cafarella,Stephen Soderland, Matthew Broadhead, and Oren Etzioni. Open information extraction from the web. In IJCAI, volume 7, pages 2670– 2676, 2007.
[25] Anthony Fader, Stephen Soderland, andOren Etzioni. Identifying relations for open information extraction. InProceedings of the Conference on Empirical Methods in Natural Language Processing, pages 1535–1545. Association for Computational Linguistics, 2011.
[26] Michael Schmitz, Robert Bart, Stephen Soderland, Oren Etzioni, et al. Open language learning for information extraction. In Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pages523–534. Association for Computational Linguistics, 2012.
[27] Andrew Carlson, Justin Betteridge,Bryan Kisiel, Burr Settles, Estevam R Hruschka Jr, and Tom M Mitchell. Towardan architecture for never-ending language learning. In AAAI, volume 5, page 3,2010.
[28] Ndapandula Nakashole, Gerhard Weikum,and Fabian Suchanek. Patty: a taxonomy of relational patterns with semantictypes. In Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pages1135–1145. Association for Computational Linguistics, 2012.
[29] T. Mitchell, W. Cohen, E. Hruschka, P.Talukdar, J. Betteridge, A. Carlson, B. Dalvi, M. Gardner, B. Kisiel, J.Krishnamurthy, N. Lao, K. Mazaitis, T. Mohamed, N. Nakashole, E. Platanios, A.Ritter, M. Samadi, B. Settles, R. Wang, D. Wijaya, A. Gupta, X. Chen, A.Saparov, M. Greaves, J. Welling. In Proceedings of the Conference on Artificial Intelligence (AAAI), 2015.

本期责任编辑：赵森栋
本期编辑：吴洋

- 精彩内容，记得分享到朋友圈 -

哈尔滨工业大学社会计算与信息检索研究中心

理解语言，认知社会

以中文技术，助民族复兴

长按二维码，关注哈工大SCIR

微信号：HIT_SCIR