北京大学学报(自然科学版)
Acta Scientiarum Naturalium Universitatis Pekinensis
doi: 10.13209/j.0479-8023.2015.036

多策略同义词获取方法研究

宋文杰[1,2]　顾彦慧[1,2,†]　周俊生[1,2]　孙玉杰[1,2]　严杰[1]　曲维光[1,2,3]

1.南京师范大学计算机科学与技术学院，南京 210023; 2.江苏省信息安全保密技术工程研究中心，南京 210023;　3.南京大学计算机软件新技术国家重点实验室，南京 210023;
†通信作者, E-mail: gu@njnu.edu.cn

摘要　提出一种多策略同义词获取方法，一方面利用《同义词词林》、《中文概念词典》等现有语义词典中蕴含的同义关系获取同义词，另一方面根据百度百科信息框(Bdbk)中特征词和汉典网(Zdic)中 HTML 标记获取同义词，同时利用 DIPRE 自动获取模式的方法，从百科文本中发现置信度较高的模式和同义关系。实验结果表明，本文的方法在 NLP&CC 2012 同义词评测数据集中取得较好结果。利用上述多策略同义词获取方法，以《现代汉语语法信息词典》名词部分为目标，构建一部同义词词典并进行人工校对，为《现代汉语语法信息词典》构建较为完善的语义关系体系作出尝试。

关键词　同义词；关系抽取；模式匹配；网络百科
中图分类号　TN914

Multi-strategies Extraction of Chinese Synonyms

SONG Wenjie[1,2], GU Yanhui[1,2,†], ZHOU Junsheng[1,2], SUN Yujie[1,2], YAN Jie[1], QU Weiguang[1,2,3]

1. School of Computer Science and Technology, Nanjing Normal University, Nanjing 210023;
2. Jiangsu Research Center of Information Security & Privacy Technology, Nanjing 210023;
3. State Key Lab for Novel Software Technology,NanjingUniversity, Nanjing 210023;
† Corresponding author, E-mail: gu@njnu.edu.cn

Abstract　Cilin and Chinese Concept Dictionary are used as dictionary resources in many NLP applications. The authors study some strategies on Chinese synonyms extraction according to key word of the infobox in baidubaike and HTML tag of the web page in Zdic. Meanwhile, DIPRE (Dual Iterative Pattern Relation Expansion) is applied to discover high credible patterns and synonymous instances in Encyclopedia corpora. Extensive experimental evaluation demonstrates that proposed strategies outperform the NLP&CC 2012 evaluation results. A sophisticated synonym dictionary is built with manually proofreading for Noun part of The Grammatical Knowledge-base of Contemporary Chinese, which would make contributions to perfecting the semantic systems of the Grammatical Knowledge-base of Contemporary Chinese.

Key words　synonym; relation extraction; pattern-based method; Encyclopedia

同义关系是非常重要的语义关系，常被运用于词义消歧[1]、信息检索[2]、文本分类[3]等自然语言处理任务。现有的同义词词典主要有《同义词词林》(扩展版)(Cilin)和《中文概念词典》[4](Chinese Concept Dictionary, CCD)等，其中《同义词词林》原版由梅家驹[5]1983 年编著完成，哈工大对《同义词词林》进行更新并发布了 Cilin，将原有的 3 层体系扩充到 5 层并删除掉一些罕见词语。这些传统的语义词典往往存在收录规模有限，新的复合词、专有名词(例如机构名、地名)等在词典中覆盖

国家自然科学基金(61272221，61472191)、国家社会科学基金(11CYY030，10CYY021)、江苏省社会科学基金(12YYA002)和江苏省高校自然科学基金(14KJB520022)资助
收稿日期：2014–06–30；修回日期：2014–10–27；网络出版时间：2014-12-01 09:26

率低，词条更新缓慢等缺陷。

传统的语义词典需要与时俱进地进行更新。随着时代的发展，网络百科和其他网络资源可以提供的信息越来越多，其中网络百科包含大量的专有词汇的描述，这些描述不仅结构清晰、内容丰富而且拥有大规模领域志愿者定期更新维护，而其它网络资源也可以提供丰富的语义信息，合理利用这些资源可以弥补传统同义词词典的缺陷。基于上述考虑，本文提出一种多策略同义词获取方法，以现有同义词词典资源为基础，利用百度百科信息框、汉典网 HTML 标记抽取同义词，同时利用 DIPRE(Dual Iterative Pattern Relation Expansion) [6] 自动发现百科文本中的新模式从而获取同义词。实验证明，与 NLP&CC 2012 同义词评测第一名相比，本文提出的方法取得了更好的结果。利用上述方法，以《现代汉语语法信息词典》[7] (The Grammatical Knowledge-base of Contemporary Chinese, GKB)名词部分为目标构建了同义词词典(以下简称 GKB_Synonym)，并对其进行人工校对，为完善 GKB 的语义体系进行了初步尝试。

1 相关工作

传统的语义关系抽取主要包括基于模式匹配和基于统计的方法。基于模式匹配的方法运用人工或自动抽取的规则发现词汇语义关系，Hearst[8]最早发现两个名词之间特定的词汇句法模式预示着特定的语义关系。基于统计的方法根据统计词语和相近词语联合出现的概率，来判断两个词语是否有语义关系。Collins 等[9]对语句进行句法分析，将句法分析树切分成子树，并将句法分析树的子树作为特征构建特征向量，利用感知器算法对英文句子进行语义关系抽取。基于模式的方法从语料库中直接抽取语义关系，准确率较高，但是模式的适用性和语料本身密切相关，移植到新领域成本比较高。基于统计的方法适合处理大规模数据，移植性好、易扩展，但是准确性要比基于模式的抽取方法低。

NLP&CC 2012 将中文词汇语义关系抽取任务作为会议评测的一部分，主要针对同义和上下位关系抽取发布样例数据集和评测数据集，这也是国内第一个开放的、相对标准的语义关系抽取数据集。NLP&CC 2012 同义词评测中取得较好成绩的几只队伍并未单一使用传统的基于模式[10]的方法，而是利用现有的词典资源和网络资源[11]，考虑多种策略相结合的方法[12]。现有语义词典资源方面：孙玉霞等[13]、范庆虎[14]利用 CCD 和 Cilin 获取同义词集合，刘燚灵等[15]将哈工大义典作为系统使用的结构化同义词典。网络资源方面：孙玉霞等在百度百科中使用模式匹配和并列结构的方法，提出一系列过滤规则对候选同义词进行过滤；范庆虎等根据 34 个特征词和标点符号，定位目标词在百科文本中的上下文，然后根据标点符号抽取同义词，同时利用有道在线翻译，抽取目标词英文对应中文翻译网页中同义词；刘燚灵等利用百度百科和豆瓣中带有结构标记的网页，构造 HTML 模板，挖掘词汇同义关系。

上述方法存在一些不足，孙玉霞等利用基于词汇字面相似度的方法进行过滤，只考虑了字面结构，忽略了语义方面的因素。范庆虎等采用有道在线翻译的方法，在英汉互译的过程中出现的错误传播会获取噪音词语，导致准确率降低。刘燚灵等并未使用 Cilin 和 CCD 作为同义词词典资源，而实验证明这两部词典对于同义词获取十分重要。

2 多策略同义词获取方法

2.1 基于词典的获取方法

基于词典的获取方法主要从现有的同义词典中获取同义关系。Cilin 和 CCD 两部词典都基于语义类组织，其中 Cilin 将近 7.8 万个词语，而 CCD 收录概念总数近 10 万。从这些人工编撰的词典中获取同义词准确率较高，一方面缓解了同义词获取的困难，另一方面也为后续基于模式的同义词获取提供了种子实例。

Cilin 中同义词即为包含目标词的、编码末尾为“=”的词语集合，例如 Cilin 中“Ae07C01=渔民 渔翁 渔家渔夫 渔父 打鱼郎”，根据“=”可以得到“渔民 渔翁 渔家 渔夫 渔父 打鱼郎”构成的同义词集合。CCD 中概念采用同义词集(CSynset)表示，获取 CCD 中目标词的同义词只需要查询 CCD 数据库中一行记录的“CSynset”字段是否包含该词语，但是 CCD 中一个词语可能出现在多个概念中，需要对候选同义词集合去重。表 1 为“兴趣”在 CCD 中的部分概念集合，“Offset”表示该行在 CCD 数据库中的行号，“CDefinition”是概念的中文定义。根据观察可知，目标词在不同概念中的同义词集合并不完全是严格的同义关系，而有可能只是某种意义上

表 1　“兴趣”对应的概念集合(部分)
Table 1　Part of the synsets of “interest”

Offset	CDefinition	CSynset
00273579	一种主题或追求占据一个人的时间和思想	兴味、兴致、兴趣、意兴、趣味
00274273	一种你喜欢或你出众的行为	兴趣、命运、强项、绝技
04006407	想做某事的理由	兴味、兴致、兴趣、意兴、理由、缘故、趣味
04042993	使人发生兴趣的力量	兴味、兴致、兴趣、情趣、意兴、趣味

的语义相关, 因此本文根据孙玉霞[12]提出的典型同义词过滤方法, 对抽取到的同义词集合进行过滤, 过滤后的到的同义词集合为“兴味、兴致、意兴、趣味、情趣”, 过滤掉“强项、绝技、理由、缘故”等词语。

2.2　基于网页标签的获取方法

基于网页标签的方法主要利用网页中固定的网页标签直接获取同义词, 无需百科语料, 也无需分析句法模式。该方法获取同义词的规模和质量依赖于网络资源的质量, 本文主要采用百度百科信息框(Bdbk)和汉典网(Zdic)作为网络资源。其中, 百度百科中对生物名、机构名、地名等专有词汇都会给出详细的描述, 这些描述格式化地存在于百度百科信息框中。

如图 1 所示, “百合”在百度百科信息框中的同义词为“强瞿、番韭、山丹、倒仙”, 这些同义词出现在特征词“别称”之后。例如: 其它译名、其他叫法、中文名称、中文学名、处方用名、古代称谓、近义成语、公司名称、中文队名、中文别名、其他名称、中医学名、通用名、同义词、中文名、口语、别名、俗名、异名、姓名、名称等词语也属于特征词, 这些词均经过人工确认, 保证获取到同义关系的准确性。因此, 利用百度百科信息框进行同义词获取时, 只需要得到词汇的百科网页, 然后判断是否包含百科信息框, 再根据特征词出现的位置, 利用正则式匹配处理该段网页源码, 最终获取同义词。

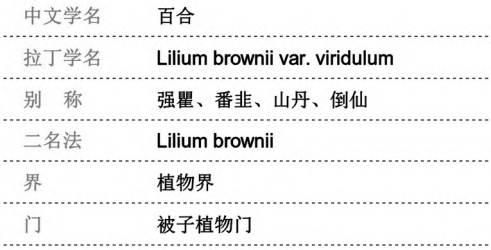

图 1　词语“百合”的百度百科信息框(部分)
Fig. 1　Part of infobox of Lily in Baidubaike

汉典网成立于 2004 年, 是一个开源的在线词典网站。如图 2 所示, 在“兴趣”的搜索结果页面中会直接给出其同义词集合。在利用汉典网获取同义词时, 首先获取“同义词”标签对应的 HTML 标记, 然后利用正则式匹配获取同义词。例如下面一段网页代码:

<imgsrc="/images/c_i_tyc.gif" align="absmiddle">风趣趣味有趣兴致…… 。

代码中第一行“ <imgsrc="/images/c_i_tyc.gif" align="absmiddle">”和最后一行“ ”就是汉典网的HTML标记, 该标记是格式化的, 利用正则式匹配处理标记间的网页信息, 可以获取到“风趣”、“趣味”、“有趣”和“兴致”等同义词。

2.3　基于模式的获取方法

在关系抽取领域, DIPRE 利用模式和关系之间的二元性(duality), 使用自举(bootstrapping)的模式匹配方法, 在自然语言文本上抽取关系实例。在迭代过程中, 抽取出的置信度较高的关系实例添加到正例集合, 产生的新的模式集合在下一次迭代中得以应用。该方法对信息冗余[16]的依赖较大, 通常在一个相对较小的数据集上, 信息冗余的假设通常弱化, 甚至不成立。这样会导致一些种子关系实例, 或中间过程中抽取的关系实例不一定能产生足够多样化的模式, 最终影响新模式和实例对的获取。Pantel 等[17]针对 DIPRE 的缺陷提出 Espresso 方法, 主要利用网络中海量的文本信息, 在整个 Web 上挖掘大量的文本模式。本文基于 Pantel 的思想, 利用网络百科爬取词条的百科语料, 然后计算模式和

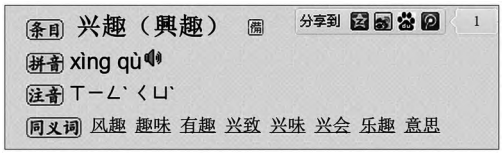

图 2　词语“兴趣”的汉典网页(部分)
Fig. 2　Part of web page of “interest” in Zdic

关系实例的可信度, 具体步骤如下。

1) 从 Cilin 和 CCD 中获取的同义关系中抽取同义关系对 $i=\{x, y\}$, 人工校对后将部分同义词对作为种子实例加入集合 I。

2) 基于 Web 爬取百科语料, 在百科语料中搜索包含集合 I 中同义实例对 i 的句子, 然后获取实例对 i 中词语 x 和 y 之间的新模式 p, 由式(1)可得模式 p 的可信度 $r(p)$:

$$r(P)=\frac{\sum_{i\in I}\frac{\mathrm{pmi}(i,p)}{\max_{\mathrm{pmi}}}\times r_l(i)}{|I|}, \tag{1}$$

其中, $r_l(i)$表示实例对 i 的可信度, 人工定义的种子实例的 $r_l(i)$设为 1, $\max_{\mathrm{pmi}}$为所有模式和实例对之间互信息的最大值, $r_l(p)$的取值范围为[0,1]。式(1)中, 实例 $i=\{x, y\}$和模式 p 的互信息 $\mathrm{pmi}(i, p)$由式(2)可得:

$$\mathrm{pmi}(i,p)=\log\frac{|x,p,y|}{|x,*,y||*,p,*|} \tag{2}$$

其中$|x, p, y|$为实例对 i 和模式 p 共同出现的次数, * 为通配符。当 $r(p)\geqslant\alpha$ 时, 模式 p 认为是可信模式。

3) 利用式(2)中可信模式 p 获取百科文本中新的同义词对, 同义词对的可信度定义为$r_l(i)$, 由式(3)可得

$$r_l(i)=\frac{\sum_{p\in P'}\frac{\mathrm{pmi}(i,p)}{\max_{\mathrm{pmi}}}\times r(p)}{|P|}, \tag{3}$$

其中 $r(p)$为模式 p 的可信度, 根据实验可知, 当 $r_l(i)\geqslant\beta$ 时, 获取的新实例对为可信同义关系。

4) 重复迭代步骤 2 和 3, 直到百科文本中没有新的模式或新的同义词被获取。

上述 α 和 β 的取值可以根据准确率和召回率之间的取舍进行调整。例如, 人工构建同义词实例对(氢氧化钠, 烧碱), 然后在百科文本语料中匹配实例对, 发现句子“氢氧化钠通常称为烧碱。”, 从而获取新候选模式 p “通常称为”。利用获取到的新模式 p, 在语料中匹配发现句子“移动电话通常称为手机”, 从而获得候选同义关系对(移动电话, 手机)。表 2 给出部分模式、语料中模式所在句子和提取到的同义词对。

3 实验

3.1 同义词评测

NLP&CC 2012 语义评测发布的同义词测试集中包含 778 个词语, 采用评测中的 3 个指标[17]: 准确率(P), 召回率(R) 和 F 值(F-measure), 分别计算宏平均和微平均。利用基于 Cilin, CCD, 汉典(Zdic), 百度百科信息框(Bdbk)和模式(Pattern)的方法抽取同义词, 表 3 为将目标词在不同方法中的同义词集合合并、去重后作为测试集结果的微平均和宏平均的准确率 P、召回率 R 和 F 值。由表 3 可知, 基于 Cilin 和 CCD 方法的微平均召回率达到 55.12%, 宏平均召回率达到 52.32%, 构成 GKB_Synonym 的基础, 直观地体现出 Cilin 和 CCD 在同义词获取中的重要性。

表 4 中郑州大学、华为、南京师范大学为 NLP&CC 2012 评测前三名, 本文方法为将本文第 1 节中提出的多策略获取结果进行合并得到的同义词词典, 与 NLP&CC 2012 评测第一(南京师范大学)相比, 微平均 F 值达到 53.54%, 提高了 12.54%,

表 2 由 DIPRE 发现的模式和实例对

Table 2 Patterns and synonyms discovered by DIPRE

模式	语料句子	实例对
又称	公休假日又称“公休日”	(公休假日, 公休日)
即	师大即师范大学	(师大, 师范大学)
也称	牛王节也称“牛神节”	(牛王节, 牛神节)

表 3 逐步合并各个方法后试验结果

Table 3 Result after gradually merging different strategies to extract synonyms

抽取策略	宏平均			微平均		
	准确率/%	召回率/%	F 值/%	准确率/%	召回率/%	F 值/%
Cilin+CCD	38.91	52.32	39.06	41.01	55.12	47.03
+Zdic	30.21	43.02	43.02	42.11	60.07	49.32
+Bdbk	37.00	50.30	38.20	43.21	64.23	51.65
+Pattern	43.11	57.53	49.29	44.31	67.63	53.54

表 4　NLP&CC 2012 同义词评测数据集实验结果对比
Table 4　Comparison of different startegies in NLP&CC 2012 synonyms evaluation data set

抽取策略	宏平均			微平均		
	准确率/%	召回率/%	*F* 值/%	准确率/%	召回率/%	*F* 值/%
郑州大学	32.56	69.30	39.19	25.40	70.40	37.34
华为	36.41	51.76	36.64	27.54	58.29	37.40
南京师范大学	35.88	60.41	39.68	30.25	63.58	41.00
本文方法	43.11	57.53	49.29	44.31	67.63	53.54

宏平均 *F* 值达到 49.29%, 提高了 9.61%。

3.2　同义词词典 GKB_Synonym

目前为止 GKB 中尚未构建同义词集合, 由于同义关系更多地在名词中出现, 而且 GKB 中名词所占比例也最高(GKB(1999 年版)[18]总库共包含 73879 个词语, 其中名词 35203 个, 占 47.65%)。因此, 本文以 GKB 名词部分为目标, 利用第 2 节中的多策略方法获取同义词, 并对获取的候选词语进行人工校对, 得到同义词词典 GKB_Synonym。人工校对步骤如下。

步骤 1 将目标词 *W* 抽取到的同义词候选集合 *S* 均分为 5 份, 每份由两人分别进行校对;

步骤 2 双人校对过程中, 将正确的同义词加入集合 *P*, 其余的词语被加入非同义词集合 *N*, 若需要对候选集合进行扩充, 将人工输入的同义词加入集合 *M*;

步骤 3 取 *P* 和 *M* 的并集作为目标词 *W* 的同义词集合 *R*, 加入到 GKB_Synonym。

本文使用的 GKB 是第 2 版, 约 8 万多词条, 其中名词部分包含 36738 个单义词。如表 5 所示, 人工校对之前共有 6684 个目标词语未抽取到同义词。经过人工校对后, 共 13597 个目标词不包含同义词, 过滤了更多非同义词集合。新词覆盖率方面, 经过校对后的 GKB_Synonym 收录了 32070 个新词, 占总库的 58.08%, 说明网络资源提供的同义词语可以较好地解决现有词典的新词覆盖率低的问题。经统计, GKB_Synonym 共包含 23141 个同义词集合和 55211 个词语, 经过人工校对后可知基于本文的多策略抽取方法准确率为 63%。

表 5　GKB_Synonym 人工校对前后对比
Table 5　Result of manualy proofreading in GKB_Synonym

GKB_Synonym	未抽取到名词数	新词数	总词数
工校对前	6684	57586	87640
人工校对后	13597	32070	55211

由于 GKB_Synonym 是以名词为目标构建, 而 Cilin 中获取的同义词集合包含名词、动词、形容词等多种词性。为了便于比较两者的差别, 本文对 Cilin 中抽取的同义词集合进行如下处理: 首先利用复旦大学提供的自然语言处理工具包 FudanNLP[19]对同义词集合进行词性标注, 若集合中 2/3 的词语都是名词, 则将被标注为名词词性的词语加入 Cilin_Noun。经统计, Cilin_Noun 共包含 3581 个同义词集合, 占 Cilin 的 35.91%, 总词数占 29.24%。本文获取的 GKB_Synonym 共包含 55211 个词语, 是 Cilin_Noun 规模的 5 倍。表 6 给出部分词语的同义词集合, 可以看出 GKB_Synonym 不仅扩充了部分 Cilin_Noun 中名词的数量, 而且加入了 Cilin_Noun 中未包含的新名词, 从规模和质量上均取得较好的结果。

4　结语

本文针对现有同义词语义资源更新缓慢、新词覆盖率低、专有名词数量少等问题提出了一种多策略同义词获取方法, 主要贡献如下:

1) 利用 DIPRE 自动发现百科文本中新的模式, 计算同义词对与新模式的可信度, 该方法可以无监

表 6　同义词集合比较
Table 6　Comparison of synsets

词语	Cilin	GKB_Synonym
百合	百合花	强瞿、番韭、山丹、倒仙、百合花
白果	银杏	银杏、鸭脚子、灵眼、佛指柑、公孙树子、银杏子、佛指甲
词典	–	辞典、辞书、字典
磁卡	–	磁条卡、磁性卡片、磁卡片、磁性卡

督地在大规模语料中获取模式和实例对获取。以开源的在线词典汉典网作为同义词获取的新途径，提高同义词获取的召回率。

2) 为 GKB 名词部分构建了同义词词典 GKB_Synonym。与现有的 Cilin 相比，不仅扩充了已有词语的规模，同时加入了部分新名词、新术语，使得 GKB_Synonym 拥有更为丰富的同义词表述。经过人工校对后的 GKB_Synonym 即可以作为测试数据集，用于同义词关系获取和验证，也可以被用于其他自然语言处理任务。

下一步工作将研究同义词获取方法，继续扩展 GKB_Synonym，并利用 GKB_Synonym 对同义关系验证进行研究。由于本文构建的同义词词典主要针对 GKB 名词部分，而且基于百度百科信息框和汉典网的方法并不适合其他词性的同义关系获取，针对其他词性的同义词获取进行研究对 GKB 全词性语义体系的构建具有重要意义。

参考文献

[1] Li Xiaobin, Szpakowicz S, Matwin S. A WordNet-based algorithm for word sense disambiguation // International Joint Conference on Artificial Intelligence. Montreal: IJCAI'95, 1995: 1368–1374

[2] 曹晶. 同义词挖掘及其在概念信息检索系统中的应用研究[D]. 沈阳: 东北大学, 2006

[3] 张剑, 李春平. 基于 WordNet 概念向量空间模型的文本分类. 计算机工程与应用, 2006, 42(4): 174–178

[4] 于江生, 刘扬, 俞士汶. 中文概念词典规格说明.汉语语言与计算学报, 2003, 13(2): 177–194

[5] 梅家驹, 竺一鸣, 高蕴琦, 等. 同义词词林. 上海: 上海辞书出版社, 1983

[6] Brin S. Extracting patterns and relations from the world wide web // The World Wide Web and Databases. Berlin: Springer, 1999: 172–183

[7] 俞士汶, 朱学锋, 王惠, 等. 现代汉语语法信息词典说明书. 中文信息学报, 1996, 10(2): 1–22

[8] Hearst M A. Automatic acquisition of hyponyms from large text corpora // Proceeding of the 14th conference on Computational Linguistics. Pennsylvania: Association for Computational Linguistics, 1992: 539–545

[9] Collins M, Duffy N. Convolution kernels for natural language // Advances in neural information processing systems. Vancouver: NIPS, 2001: 625–632

[10] 陆勇, 侯汉清. 基于模式匹配的汉语同义词自动识别. 情报学报, 2006, 6(25): 720–724

[11] Lu Yong, Hou Hanqing. Research on automatic acquiring of chinese synonyms from Wiki repository // Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and International Conference on Intelligent Agent Technology. Sydney: IEEE, 2008, 3: 287–290

[12] Lu Yong, Zhang Chengzhi, Hou Hanqing. Using multiple hybrid strategies to extract chinese synonyms from encyclopedia resource // Fourth International Conference on Innovative Computing, Information and Control. TaiWan: IEEE, 2009: 1089–1093

[13] 孙玉霞, 狄颖, 曹冉, 等. 中文同义词自动抽取研究 [EB/OL]. (2012–11–19)[2014–05–27]. http://tcci.ccf.org.cn/conference/2012/pages/page10_nlpcc2012testpaper.html

[14] 范庆虎, 眷红英, 张坤丽, 等. 基于词典和 Web 的词汇关系抽取[EB/OL]. (2012–11–19)[2014–05–27]. http://tcci.ccf.org.cn/conference/2012/pages/page10_ nlpcc2012testpaper.html

[15] 刘燚灵, 吉阳生, 顾翀, 等. 面向开放异构知识库的词汇同义关系学习[EB/OL]. (2012-11-19)[2014-05-27]. http://tcci.ccf.org.cn/conference/2012/pages/page10_nlpcc2012testpaper.html

[16] 王刚. 自动抽取维基百科文本中的语义关系[D]. 上海: 上海交通大学, 2008: 4–11

[17] Pantel P, Pennacchiotti M. Espresso: leveraging generic patterns for automatically harvesting semantic relations // Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics. Sydney: Association for Computational Linguistics 2006: 113–120

[18] 俞士汶, 朱学锋. 《现代汉语语法信息词典》的新进展. 中文信息学报, 2001, 15(1): 59–64

[19] Qiu Xipeng, Zhang Qi, Huang Xuanjing. FudanNLP: a toolkit for Chinese natural language processing // Proceedingsof Annual Meeting of the Association for Computational Linguistics. Sofia: Association for Computational Linguistics,2013: 49–54