文章编号:1004-9037(2014)05-0821-07

基于词典信息和网络百科的下位词获取

宋文杰[1,2] 周俊生[1,2] 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)

摘要:对中文下位词自动抽取方法进行研究,提出一种基于词典信息和网络百科的下位词获取方法,旨在构建一个较为完善的上下位词语知识库。基于词典信息的抽取方法利用《中文概念词典》和《中国分类主题词表》中蕴含的格式化信息获取上下位关系。基于网络百科的抽取方法利用维基百科、百度百科和互动百科,分析百科网页地址和内容格式,利用正则式抽取下位词语。对获取到的下位词进行自动过滤和人工校对,实验表明,与NLP&CC 2012 上下位关系评测结果相比,本文方法取得较好效果。

关键词:下位词;词典;网络百科;模式方法;正则式

中图分类号:TP391 **文献标志码**:A

Chinese Hyponymy Extraction Based on Dictionary and Encyclopedia Resources

Song Wenjie[1,2], *Zhou Junsheng*[1,2], *Qu Weiguang*[1,2,3]

(1. School of Computer Science and Technology, Nanjing Normal University, Nanjing, 210023, China;
2. Jiangsu Research Center of Information Security & Privacy Technology, Nanjing, 210023, China;
3. State Key Lab for Novel Software Technology, Nanjing University, Nanjing, 210023, China)

Abstract: Hyponymy, a kind of basic semantic relation between words, is widly used in areas, including text classification and information retrieval. Automatic extraction of such relation is an important issue in natural language processing. Two kinds of hyponymy extraction strategy, i.e., dictionary based strategy and encyclopedia based strategy are proposed to build a sophisticated hyponymy knowledge base. Chinese Concept Dictionary and Chinese Classied Subject Thesaurus are used as dictionary resources. Manual regex is introduced to extract hyponym from wikipedia, baidubaike and hudongbaike based on addresses of web pages. Extensive experimental evaluation demonstrates that the proposed strategies outperform the NLP&CC 2012 evaluation results.

Key words: hyponymy; dictionary; encyclopedia; pattern-based method; regular expression

引 言

词汇语义关系抽取评测是 2012 年 CCF 自然语言处理与中文计算会议(以下简称 NLP&CC 2012)的一部分,评测对象是中文词义关系抽取中的核心技术,包括同义关系和上下位关系。其中,上下位关系获取意义重大[1],这项工作可以对本体、知识库、词典进行正确性检测、扩充和完善,也是非格式化信息转化为格式化信息过程的重要步骤。上下位关系常作为知识库的基础框架[2-4]描述概念之间的语义关系,同时也被应用于文本分类[5-6]、信息抽取[7]、信息检索[8]等自然语言处理任务。目前,上下位关系的获取方法主要有两种:一种是基于模式[9-10]的上下位关系获取,主要利用模式匹配发现上下位关系;另一种是基于统计[11-13]的

基金项目:国家自然科学基金(61272221,61472191)资助项目;江苏省社会科学基金(12YYA002)资助项目;国家社会科学基金(11CYY030,10CYY021)资助项目。

收稿日期:2014-06-02;**修订日期**:2014-09-01

上下位关系获取方法，主要基于语料库和统计语言模型，计算概念之间的关联度，从而获取概念间上下位关系。基于模式的方法从语料库中直接抽取词条关系，获取的概念在概念层次上较接近，准确率比较高。但是如果没有高质量的语料库，基于模式的方法获取的有意义的匹配数不多，会导致关系获取的召回率偏低。基于统计的方法适合处理大规模数据，移植性好，但是其准确性要比基于模式的上下位抽取方法低。

NLP&CC 2012 下位词评测中取得较好结果的两支队伍并未单一地考虑传统的基于模式的方法，而是利用现有的词典资源和网络资源，考虑多种策略相结合的方法。郑州大学范庆虎等[14]利用中文概念词典中蕴含的上下位关系，并且结合百度相关搜索、百度百科标签、互动百科、维基百科等 Web 方式进行抽取，并对获取的结果进行合并和过滤。北京交通大学刘江鸣等[15]采用基于多策略并行的方法，其中包括学科分类、词汇细化、开放分类和模板匹配等，并构造上下位关系识别系统，动态地查询和构造上下位关系表。两支队伍在最终的评测结果中分别取得了宏平均和微平均 F(F-measure)值第一的成绩。但是上述两支队伍的方法仍然存在不足，例如郑州大学的方法虽然利用了网络百科，但是抽取规模均比较小，比如：百度标签只获取首页词语，互动百科页面中未使用“全部词条”对应的链接，维基百科中未考虑“本分类”下的词语，百度相关中获取的词语受到百度搜索的影响噪音较多。北京交通大学的方法在学科分类策略中仅考虑生物学领域，开放分类策略的方法只能获取目标词的上位词，模板匹配中利用词典匹配来获取模板句子中蕴含的上下位词语对未登录词的识别难度较大。

针对上述研究存在的问题，本文将提出一种基于词典知识和网络百科的下位词获取方法，旨在利用词典和网络百科资源构建一个较为完善的下位词语知识库。

1 词典信息和网络百科的下位词获取

1.1 基于词典信息的下位词获取方法

现有的词典资源十分宝贵，中文方面有《中国分类主题词表》、《同义词词林》和《中文概念词典》等，这些词典都蕴含上下位关系，在信息抽取、自然语言处理等领域起着重要作用。本文使用的词典资源主要包含《中文概念词典》和《中国分类主题词表》。

《中文概念词典》(Chinese concept bictionary，CCD)是北京大学计算语言学研究所开发的与 WordNet 兼容的汉语语义词典。CCD 中名词数据库包括 15 个字段，在抽取数据库中上下位词语关系时，只需要利用其中的 3 个字段信息：Offset，CSynset，Hyponym。其中，Offset 指概念的 ID 编号，其在 CCD 中唯一存在，用 8 位数字表示；CSynset 指中文同义词集，表示 CCD 中的一个概念；Hyponym 指下位词编号，每个指针占据 8 位，代表其“下位概念”的 ID 编号。

如表 1 所示，CCD 名词数据库第 2 876 项中 3 个字段信息为(00454943，休克疗法 震击疗法，00453177004534010045358 2)。根据定义，抽取该项中 Hyponym 字段内容中每 8 位对应的概念，可以得到第 2 877 项概念“电休克 电击疗法 电休克疗法”、第 2 878 项概念“胰岛素休克”和第 2 879 项概念“卡地阿唑休克”，对抽取出的结果进行合并、去重、去噪，就可以得到该概念对应的下位词集合。经过统计，最终抽取出的下位词词典共包含 14 031 个概念(每个概念平均由 3 个词语表示)，下位词语总数 139 918 个，每个概念的平均下位词数为 9.97。

表 1 CCD 名词数据库示例

Table 1 Samples in CCD database

Index	Offset	CSynset	Hyponym
2 876	00452943	休克疗法 震击疗法	00453177004534 0100453582
2 877	00453177	电休克电击疗法 电休克疗法	
2 878	00453401	胰岛素休克	
2 879	00453582	卡地阿唑休克	

《中国分类主题词表》(简称《词表》)于 1994 年 6 月正式出版，该表共收录 5 万余条类目，21 万余条主题词和主题标题。《词表》中主题词涵盖各领域，很好地解决了文献[15]中学科分类仅限于生物学领域的弊端。

如图 1 所示，《词表》中每一个主题词或主题标题均有固定的对应位置，在文件具体表现为主题词所在行不以空格开始。符号系统中“.”表示词族中概念的属分关系，本文主要利用《词表》文件中包含该符号的词语作为主题词的下位词。在处理《词表》文件时，首先根据主题词将文件划分为对应的信息块。图 1 中以“广播”为主题词的信息块共包

```
广播
  Broadcasting
  GSS
  ·电视广播
  ··文字电视广播
  ·卫星广播
  ·无线广播
  ··调幅广播
  ···同步广播
  ··调频广播
  ···数据广播
  ··短波广播
  ··数字广播
  ···数字音频广播
  ··中波广播
  ·有线广播
```

图 1　《词表》文件中信息块格式

Fig. 1　Text format in "Chinese Classied Subject Thesaurus"

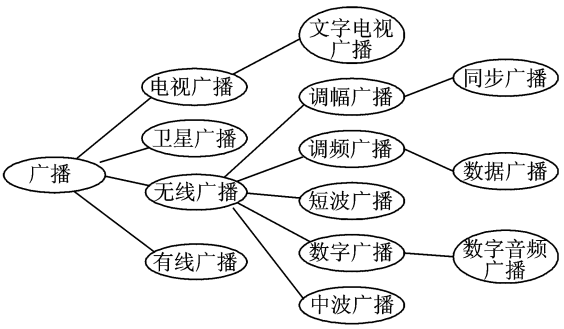

图 2　《词表》单个主题词生成的多层次树

Fig. 2　Multi-level tree generated from one subject word in "Chinese Classied Subject Thesaurus"

含 16 行文本，其中“Broadcasting”和“G22”所属行不包含属分关系符号，需要被过滤。假设“广播”在《词表》文件中的信息块为 C_i，其中包含属分关系符号的行号为 $C_{i,j}$，将根据 $C_{i,j}$ 所属层次 $L_{i,j}$ 与 $C_{i,j-1}$ 所属层次 $L_{i,j-1}$ 的关系构建以主题词为树根的多层次树，其中层次函数 L 的值由当前行包含属分关系符号“·”的个数决定，构造策略如下：

(1) $L_{i,j}=1$，$C_{i,j}$ 为多层次树中第 1 层节点，以主题词作为父节点，直接加入树中。

(2) $L_{i,j}=L_{i,j-1}$，$C_{i,j}$ 和 $C_{i,j-1}$ 为兄弟关系，$C_{i,j}$ 作为 $C_{i,j-1}$ 父节点的孩子节点加入树中。

(3) $L_{i,j}=L_{i,j-1}+1$，$C_{i,j}$ 和 $C_{i,j-1}$ 为父子关系，$C_{i,j}$ 作为 $C_{i,j-1}$ 的孩子节点加入树中。

(4) $L_{i,j}<L_{i,j-1}$，$C_{i,j}$ 和 $C_{i,j-1}$ 为隔代关系，需要计算层次差值 k，递归查找 $C_{i,j-1}$ 的第 k 代祖先节点，作为该节点的孩子节点加入树中。

下面给出以“广播”为树根的多层次树的部分构造过程：当 $j=1$，初始化树根为“广播”；当 $j=2$，$L_{i,2}=1$，满足策略(1)，将“电视广播”作为“广播”的孩子节点加入树中；当 $j=3$，$L_{i,3}=2$，满足策略(3)，将“文字电视广播”作为“电视广播”的孩子节点加入树中。根据以上构造策略，最后生成的多层次树如图 2 所示。

据此，可以由《词表》得到一部上下位词语词典，步骤如下：

(1)对《词表》资源进行预处理，得到图 1 格式的文本；

(2)逐行读取文本，以主题词为单位将文本划分成信息块，过滤不包含属分关系符号的行；

(3)根据多层次树构造策略，以主题词为树根构造多层次树，加入到森林中；

(4)遍历森林中所有多层次树，输出其中上下位关系，得到上下位词语词典。

经过统计，《词表》生成的上下位词语词典共包含 16 787 个主题词和 155 184 个下位词，每个主题词的平均下位词数为 9.24。

1.2　基于网络百科的下位词获取方法

人工构造的词典不但数量稀少而且内容更新缓慢，随着互联网的发展，维基模式和 Web2.0 技术[16]极大地推动了信息革命。本文将利用维基百科、百度百科和互动百科 3 大网络百科进行下位词的抽取。虽然百科资源不同，但是基于网页地址和内容格式的的抽取方法一致，其基本步骤如下：

(1)获取目标词百科网页地址(如表 2 所示，其中 XXX 代表目标词编码)；

(2)通过 URL 通信链接，根据不同百科网页内容的具体格式对信息进行处理，保存可用网页信息；

(3)构造正则表达式，匹配网页信息中待抽取的下位词，并用哈希表保存目标词及其下位词集合。

表 2　百科网页地址格式

Table 2　Address and encoding of encyclopedia web page

符号	地址格式	百科	编码
$U1$	http://zh.wikipedia.org/wiki/Category:XXX	维基	UTF-8
$U2$	http://baike.baidu.com/taglist?tag=XXX	百度	GBK
$U3$	http://fenlei.baike.com/XXX/list/	互动	UTF-8

第(2)步是基于网络百科抽取方法的核心步骤，郑州大学在此步骤的处理比较简单，仅抽取百科网页中的部分信息，导致许多下位词资源未被利用。本文在此作出如下改进：

(1)维基百科　郑州大学仅考虑网页中“子分类”而未考虑“本分类”下的词语,本文将扩展抽取“本分类”下的词语。事实上,这两种标题下的词语可以被认为是下位词的两种情况:子分类和实例。例如“安全软件”在维基百科分类索引搜索页面中“子分类”下包含“加密软件”“反间谍软件”“杀毒软件”、“防火墙软件”均属于某一类安全软件,而“本分类”下的“360 安全卫士”“百度卫士”“金山卫士”等词语均则是“安全软件”的实例。此外,本文扩展考虑了维基百科中词条重定向问题。例如“朝鲜语”在维基百科中会被重定向到“韩语”,此时应该将目标词重置为重定向后的词名,重新进行抽取。

(2)百度百科　郑州大学抽取百度百科下位词时抽取规模仅为一个网页,本文将扩展抽取规模,首先获取百度百科下位词结果数,然后人为设置抽取规模。NLP&CC 2012 评测给出的标准答案中每个目标词平均对应的下位词数为 22,因此本文设置抽取百度百科下位词数为 50,保证抽取规模。

(3)互动百科　郑州大学通过互动百科分类专题中搜索目标词,抽取返回网页中“分类词条”标题下的词语作为下位词,抽取规模仅为一个网页。本文扩展抽取“全部词条”标题超链接到的网页,将目标词在互动百科中所有下位词全部抽取出来。需要注意的是,必须对抽取标签中的 UTF-8 码进行转码获取下位词,因为在互动百科网页中,如果标签词语长度过长,标签内容是不完整的。

(4)抽取效率　对网页信息的处理一般采用正则式匹配,如果先将整个网页信息保存为字符流,再从头至尾进行匹配会导致抽取效率低下。因为如果词条不存在,保存全部网页的字符流将毫无意义。为了保证大规模抽取时的效率,本文将根据不同百科网页中固有的网页标志来判断网页类型,采取对应策略,减少不必要的时间开销。

1.3　过滤规则

由于网络资源存在网页格式的不规范性和人工编撰的不确定性,需要对获取到的下位词集合进行简单的过滤,过滤规则如下:

规则 1　过滤重复词语和不含中文词语。对获取到的下位词集合进行合并、去重是必要工作。此外,NLP&CC 2012 评测给出的样例数据中不包含全英文、全数字词语,但包含存在部分英文和数字的情况。例如“计算机软件”的下位词集合中“Draw”“itunes”等词被过滤,而“计算机语言”的下位词集合中的“Java 语言”“C 语言”等词被保留。

规则 2　过滤包含停用词表中符号的词语。由于网络文本存在不规范性,抽取到的中文标签内容可能存在停用词,如“怎么”“如何”“嘛”等。

规则 3　过滤前缀词。汉语作为一种词根语,根据经验可知,大部分的下位词语都以目标词作为后缀,例如“水电厂”的下位词“长庆水电厂”“金城江水电厂”等,而“水电厂实习报告”“水电厂招聘”这类以目标词作为前缀的词语需要被过滤。本文首先对包含目标词的下位词进行分词,假设下位词长度为 L,目标词所处位置为 M,当 $M<L/2$ 时将该词作为前缀词过滤。

2　实验结果与分析

本文使用 NLP&CC2012 语义关系识别任务中的评测方法,采用 3 个指标:正确率(Precision),召回率(Recall) 和 F 值,分别计算宏平均和微平均。表 3 是 NLP&CC 2012 的下位关系评测结果,其中北京交通大学宏平均正确率达到 66.21%,微平均正确率达到 70.43%,但是其宏平均和微平均召回率均低于 50%;而郑州大学 2 宏平均召回率达到 59.88%,微平均召回率达到 50.45%,体现出网络百科资源在保证下位词抽取规模的优势。

表 3　NLP&CC 2012 下位关系评测结果　%

Table 3　NLP&CC 2012 evaluation result

评测队伍	宏平均准确率	宏平均召回率	宏平均 F	微平均准确率	微平均召回率	微平均 F
中科院声学所	24.53	8.95	11.78	78.27	12.21	21.13
北京理工大学	31.03	9.01	11.79	63.83	8.96	15.72
北京交通大学	66.21	37.80	41.89	70.43	46.42	55.96
郑州大学 1	56.03	33.21	37.42	64.92	35.18	45.63
郑州大学 2	61.19	59.88	56.05	62.33	50.45	55.76

本文以 **NLP&CC** 2012 下位词文件中的 256 个词语作为目标词进行下位词抽取。鉴于抽取结果中许多词语未在评测答案中出现,本文将对抽取的结果进行人工标注,将正确的下位词补充到 **NLP&CC** 2012 标准答案中。例如,待抽取目标词为“身份”,评测答案给出的下位词集合包含:太平绅士、个人身份、新闻工作者、民族身份、居民身份、欧洲身份、欧盟公民身份、职业身份、家奴、国籍、学

徒身份、欧洲公民身份共 12 个词语，经人工标注补充后新加入 10 个下位词语：单身皇族、贵族阶级、内奸、澳门人、清倌人、庶出、小人物、贵族身份、公民身份、领导身份。

对 256 个目标词采取双人交叉标注，其步骤如下：

(1)将目标词 W 抽取到的结果 S 去除标准答案，得到 S'；

(2)由双人对 S' 进行人工标注，标注时将认为正确的下位词加入下位词集合 P，其余的自动被加入非下位词集合 N，对于目标词 W，取 P 的交集，加入到 W 的标准答案中；

(3)将 S' 中去除 P 的交集和 N 的交集得到 S''，由第三人对 S'' 进行标注，将认为正确的下位词加入标准答案，最终得到补充后的标准答案。

由于评测答案中未考虑不含中文的词语，因此在进行人工标注时，全英文和全数字词语不予考虑。经过统计，评测给出的 256 个目标词的下位词总数为 5 691，平均下位词数为 22.23。加入人工标注后的下位词总数为 12 027，平均下位词数为 46.98，比标准答案规模扩大一倍多。这不仅符合时间推移规律，也有利于本文下位词抽取效果的正确评价。

如表 4 所示，$R1$，$R2$，$R3$ 顺序对应了 1.3 节中的 3 个过滤规则。经过规则 1 过滤后，实验结果的宏平均召回率达到了 67.52%，微平均召回率达到了 77.82%，总的抽取规模超过了标准答案规模。虽然加上规则 2 和规则 3 之后召回率均有所下降，但是正确率和 **F** 值均有所提高，其中宏 **F** 值达到 55.06%，微 **F** 值达到 60.23%。

以本文实验中标准答案评价 **NLP&CC** 2012 评测结果时，进行如下假设：参加 **NLP&CC** 2012 评测队伍的抽取结果中被判定为错误的词语在本文补充后标准答案中仍然被判定为错误词语。那么，以词语为统计单位的微平均值可以重新计算。例如，北京交通大学在微平均下的召回率为 46.42%，经统计评测答案共 5 691 个词语，则正确词语为 2 641.76 个，如果以本文实验中 12 027 个词语为标准答案，则微平均召回率为 21.97%，以 70.43%作为微平均正确率，则微平均 **F** 值为 33.49%；同理可计算得到郑州大学 2 的正确词语数为 2 871.11 个，微平均召回率为 23.87%，微平均 **F** 值为 34.5%。上述数据表明，在相同评价体系下，本文的方法在微平均召回率和微平均 **F** 值上要明显好于 **NLP&CC** 2012 评测中成绩较好的两只队伍。

表 4　下位词抽取实验结果对比　%

Table 4　Comparison of different hyponyms extraction result

实验对象	宏平均准确率	宏平均召回率	宏平均 F	微平均准确率	微平均召回率	微平均 F
NLP&CC 2012 评测答案	100.00	60.37	72.34	100.00	48.92	65.70
多策略抽取($R1$)	50.88	67.52	53.16	48.10	77.82	59.45
多策略抽取($R1+R2+R3$)	57.87	63.63	55.06	50.55	74.49	60.23
北京交通大学	66.21			70.43	21.97	33.49
郑州大学 2	61.19			62.33	23.87	34.50

3　结束语

下位词抽取在自然语言处理和信息检索领域中具有重要意义。本文主要对中文下位词自动抽取的多种资源进行整合，充分利用各种词典知识和网络百科资源，旨在构建一个较为完善的上下位词语知识库。与 NLP&CC 2012 评测中取得较好结果的两支参赛队伍采用的方法相比，本文加入了《中国分类主题词表》资源，扩展了网络百科抽取规模，实验表明本文提出的方法在微平均召回率和 F 值上取得了不错的效果。

《现代汉语语法信息词典》[17-18] 是由北京大学计算语言学研究所俞士汶教授历时 10 余年研制而成的一部电子词典，其提供的信息可以广泛应用到汉语自动分析和生成、汉字识别后校正、语料库自动标注等领域，在中文信息处理领域产生广泛影响，希望利用本文的方法为《现代汉语语法信息词典》建立较为完善的上下位词体系。截止目前为止，已经完成其名词部分(共计 36 965 个)的下位词抽取工作，共获取候选下位词 445 694 个。

参考文献：

[1]　刘磊. 概念和上下位关系的获取理论和方法研究[D]. 北京：中国科学院计算技术研究所，2007：3-13.

Liu Lei. Research of theories and methods of extracting concept and hyponymy relations[D]. Beijing: Institute of Computing Technology, Chinese Academy of Science, 2007: 3-13.

[2] 张强.《中国分类主题词表》的结构及功能评介[J]. 中国图书馆学报，1995(5)：70-75.
Zhang Qiang. An evaluation and recommendation of the structure and functions of the "Chinese Classied Subject Thesaurus"[J]. Journal of Library Science In China，1995(5)：70-75.

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

[4] 于江生，刘扬，俞士汶.《中文概念词典》规格说明[J]. 汉语语言与计算学报，2003，13(2)：177-194.
Yu Jiangsheng，Liu Yang，Yu shiwen. The structure of "Chinese Concept Dictionary" [J]. Journal of Chinese Language and Computing，2003，13(2)：177-194.

[5] Wang Pu，Hu Jian，Zeng Huajun，et al. Improving text classification by using encyclopedia knowledge [C]//The 7th IEEE International Conference on Data Mining. Nebraska，United States：IEEE，2007：332-341.

[6] 王盛，樊兴华，陈现麟. 利用上下位关系的中文短文本分类[J]. 计算机应用，2010，30(3)：603-606.
Wang Sheng，Fan Xinghua，Chen Xianlin. Chinese short text classification based on hyponymy relation [J]. Journal of Computer Applications，2010，30(3)：603-606.

[7] Grishman R. Structural linguistics and unsupervised information extraction[C]//The Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction. Pennsylvania，United States：Association for Computational Linguistics，2012：57-61.

[8] Fernández M，Cantador I，López V，et al. Semantically enhanced information retrieval：an ontology-based approach[J]. Web Semantics：Science，Services and Agents on the World Wide Web，2011，9(4)：434-452.

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

[10] 刘磊，曹存根，王海涛，等. 一种基于“是一个”模式的下位概念获取方法[J]. 计算机科学，2006，33(9)：146-151.
Liu Lei，Cao Cungen，Wang Haitao，et al. A method of hyponym acquisition based on "isa" pattern[J]. Computer Science，2006，33(9)：146-151.

[11] Caraballo S A. Automatic construction of a hypernym-labeled noun hierarchy from text[C]//Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics，Association for Computational Linguistics. Pennsylvania，United States：Association for Computational Linguistics，1999：120-126.

[12] Tian Fang，Fu Jiren. Hyponymy acquisition from Chinese text by SVM[C]//Natural Language Processing and Knowledge Engineering，International Conference on IEEE. Dalian，China：IEEE，2009：1-6.

[13] Singh L，Scheuermann P，Chen B. Generating association rules from semi-structured documents using an extended concept hierarchy[C]//Proceedings of the Sixth International Conference on Information and Knowledge Management. New York，United States：ACM，1997：193-200.

[14] 范庆虎，咎红英，张坤丽，等. 基于词典和 Web 的词汇关系抽取[EB/OL]. http://tcci.ccf.org.cn/conference/2012/pages/page10_nlpcc2012testpaper.html，2012-11-19/2013-12-19.
Fan Qinghu，Zan Hongying，Zhang Kunli，et al. Lexical reiteration extraction based on dictionary and web[EB/OL]. http://tcci.ccf.org.cn/conference/2012/pages/page10_nlpcc2012testpaper.html，2012-11-19/2013-12-19.

[15] 刘江鸣，徐金安，吴培昊，等. 基于网络资源的词语语义关系自动获取[EB/OL]. http://tcci.ccf.org.cn/conference/2012/pages/page10_nlpcc2012 testpaper. html，2012-11-19/2013-12-19.
Liu Jiangming，Xu Jin′an，Wu Peihao，et al. Automatic acquisition of lexical semantic relationship based on web resource[EB/OL]. http://tcci.ccf.org.cn/conference/2012/pages/page10_nlpcc2012testpaper.html，2012-11-19/2013-12-19.

[16] 涂新辉，张红春，周琨峰，等. 中文维基百科的结构化信息抽取及词语相关度计算[J]. 中文信息学报，2012，26(3)：109-115.
Tu Xinhui，Zhang Hongchun，Zhou Kunfeng，et al. Extracting structured information from Chinese wikipedia and measuring relatedness between words[J]. Journal of Chinese Information Processing，2012，26(3)：109-115.

[17] 俞士汶，朱学锋，王惠，等. 现代汉语语法信息词典规格说明书[J]. 中文信息学报，1996，10(2)：1-22.
Yu Shiwen，Zhu Xuefeng，WangHui，et al. The

structure of the grammatical knowledge-base of contemporary chinese[J]. Journal of Chinese Information Processing, 1996, 10(2): 1-22.

[18] 冯志伟，曹右琦. 评《现代汉语语法信息词典详解》[J]. 中文信息学报，1999(2)：64.
Feng Zhiwei, Cao Youqi. Comment on "The Grammatical Knowledge-base of Contemporary Chinese" [J]. Journal of Chinese Information Processing, 1999(2): 64.

作者简介：宋文杰（1990-），男，硕士研究生，研究方向：自然语言处理、信息抽取，E-mail：njnusongwenjie@163.com；周俊生(1972-)(通讯作者)，男，副教授，研究方向：自然语言处理、信息抽取、机器学习，E-mail：zhoujs@njnu.edu.cn；曲维光(1964-)，男，教授，研究方向：自然语言处理、计算语言学、语言工程、人工智能。