ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 32,No. 1,Jan. 2017,pp. 175－181
DOI:10.16337/j.1004-9037.2017.01.021

http://sjcj.nuaa.edu.cn
E-mail:sjcj@nuaa.edu.cn
Tel/Fax:＋86-025-84892742

融合语义类信息的句法分析统计模型

袁里驰

(江西财经大学信息管理学院,南昌,330013)

摘　要:稀疏数据严重影响句子结构分析模型的结果,而句法结构是语义内容和句法分析形式的结合。本文在语义结构信息标注的基础上提出了一种基于语义搭配关系的词聚类模型和算法,建立基于语义类的头驱动句子结构分析统计模型。该语言模型不但比较成功地解决了数据稀疏问题,而且句子结构分析系统性能也有了明显的提高。句子结构分析实验结果表明,基于语义类的头驱动的句子结构分析统计模型,其召回率和精确率的值相应为88.26%和88.73%,综合指标改进了8.39%。

关键词:句子结构分析统计模型;语义角色标注;词的自动聚类;头驱动

中图分类号:TP391　　**文献标志码**:A

Statistical Syntactic Parsing Model Fusing Semantic Category Information

Yuan Lichi

(School of Information Technology, Jiangxi University of Finance and Economics, Nanchang,330013,China)

Abstract: Data sparseness severely affects the system performances of syntactic parsing, and syntactic structures are unities of syntactic forms and semantic contents. Based on the labeling of semantic information, a word clustering model and algorithm is proposed. And a head-driven statistical syntactic parsing model based on semantic category is established. The problem of data sparseness is successfully solved, and the system performances of syntactic parsing are obviously enhanced. Experiments are conducted for the head-driven statistical syntactic parsing model based on semantic category. It achieves 88.73% precision and 88.26% recall. F measure is improved 8.39% compared with the distinctive head-driven parsing model.

Key words: statistical syntactic parsing model; semantic role labeling; word clustering; head-drive

引　　言

句子结构分析是自然语言处理的一个最基本的问题,同时也是自然语言处理的关键技术之一。句子结构分析的主要目标是依据一定的句法规则分析出句子的短语组成关系,即句子包括的短语以及短语组成之间的语义、语法联系。主要的句子结构分析方法分为两种途径:基于统计的句子结构分析方法[1-11]和基于规则的句子结构分析方法。当前,句子结构分析方法主要有依存分析方法和短语组成分

基金项目:国家自然科学基金(61262035,61562034)资助项目;江西省自然科学基金(20142BAB207028)资助项目;江西省科技支撑计划(20151BBE50082)资助项目。
收稿日期:2015-05-16;**修订日期**:2015-06-02

析方法。句子短语组成分析方法主要基于上下文无关概率文法(Probabilistic context free grammar, PCFG)。早期的句子结构分析上下文无关模型从标注句法树库中直接抽取语法规则,并且将相对出现次数计算为语法规则的概率[12]。这种句法分析模型实现容易,然而以前的句法分析研究证明这类句法分析模型的效果并不令人满意,其重要原因是:上下文无关概率语法里的一些独立假定在实际中可能并不正确。依存关系文法[13-15]容易标注、结构简单,渐渐得到重用。虽然目前汉语依存关系语法分析研究取得了一定的进展,但是其准确率和效率仍然不能满足实际应用的需要。Collins[11]等学者将词汇的依存关系引入到语法中,提出了一种词汇化的上下文无关概率句子结构分析方法,推动了句子结构分析技术和方法的飞速发展。该方法的基本思想就是将短语中心词和词汇等语义信息融入上下文无关语法规则,此两类语义信息的融入,大大提高了句子结构分析方法的消歧效果,但该方法产生了比较严重的稀疏数据难题。句子结构分析是语义分析和短语结构分析的有机结合。句法分析不仅需做短语结构分析,比如句子主要短语组成分析、句子型式分析及短语成分联系分析等等,并且还必须做相关的语义联系分析。对语义联系分析越深刻和全面,将更能够对短语结构上的种种语言问题给予合理和科学的解答。词汇句子结构分析的当前模型如依存关系语法、头驱动的句子结构分析方法[11]只引入词语的语义依存信息,但没有考虑语义方面其他有关信息,比如词语语义搭配、词语的语义类等语义有关知识,然而一些语义有关的知识对语义关系、句子结构的计算和分析非常有用。语义关系分析是自然语言理解的一个关键技术问题。作为当前的自然语言研究热点课题之一,语义角色的标注[16-19](Semantic role labeling, SRL)是浅层语义关系分析的一种。语义角色标注是在句子成分级别进行浅层的语义关系分析,即对于给定的一个句子,对该句中的每个谓词成分标注出对应的语义关系成分,并且确定其对应的语义关系标记,如施事成分、受事成分、工具成分或附加语成分等。当前的句子结构分析方法还不能够成功地描述出中文语言的基本特点[20-23],使得当前中文语义关系、句子结构的计算和分析的结果相比英语差距很明显。针对传统句法结构分析统计方法存在的一些问题,本文建立了一种新颖的融合词语语义类信息的句法结构分析模型,提出了一种基于词语语义搭配关系的词聚类模型和相应算法,解决句法结构分析统计模型在引入词汇信息时带来的稀疏数据问题。

1 基于语义相似度的词聚类模型和算法

词汇化句子结构分析模型如头驱动句子结构分析方法,为了利用语义知识,句子语法生成式中的任何一个非终结符号均引入词性/核心词等语义知识。然而语义知识的引进产生了稀疏数据难题。建立基于语义类的词类语言模型[24-28]替换基于词的语言模型是缓解句子结构分析方法稀疏数据难题的主要途径之一。依据词语的语法特点和词语语义搭配之间的联系对词聚类极为重要。虽然语言学家可根据所掌握的语言信息对词分类,然而结合语言信息,应用统计方法自动分类词的办法应该更为可行。

1.1 词的聚类模型

假定 w_1, w_2 是含有语义搭配联系 Rel 的二元词组,本文用三元数组 (w_1, rel, w_2) 代表二元词组及两个词之间的语义联系。二元词组 (w_1, w_2) 在语义联系 rel 下的点互信息可定义为

$$I_{rel}(w_1, w_2) = \log \frac{p(w_1, w_2 \mid rel)}{p(w_1 \mid rel)\, p(w_2 \mid rel)} \tag{1}$$

其中

$$p(w_1, w_2 \mid rel) = \frac{p(w_1, rel, w_2)}{p(rel)}$$

这里的概率计算使用极大似然估计方法计算如下,即

$$p(w_1, rel, w_2) = \frac{\mathrm{Count}(w_1, rel, w_2)}{\mathrm{Count}(*, *, *)} \tag{2-a}$$

$$p(w_1 \mid rel)=\frac{\mathrm{Count}(w_1,rel,*)}{\mathrm{Count}(*,rel,*)} \tag{2-b}$$

$$p(w_2 \mid rel)=\frac{\mathrm{Count}(*,rel,w_2)}{\mathrm{Count}(*,rel,*)} \tag{2-c}$$

$$p(rel)=\frac{\mathrm{Count}(*,rel,*)}{\mathrm{Count}(*,*,*)} \tag{2-d}$$

其中 $*$ 表示可能的词或语义联系,因而有

$$I_{rel}(w_1,w_2)=\log\frac{\mathrm{Count}(w_1,rel,w_2)\mathrm{Count}(*,rel,*)}{\mathrm{Count}(w_1,rel,*)p(*,rel,w_2)} \tag{3}$$

定义 1 二元词组 w_1, w_2 在语义联系 rel 下的近似度由式(4,5)定义

$$\mathrm{sim}_{rel}(w_1,w_2)=\sum_{w}p(w)\frac{\alpha_{rel}}{\alpha_{rel}+\mid I_{rel}(w_1,w)-I_{rel}(w_2,w)\mid} \tag{4}$$

$$\mathrm{sim}_{rel}(w_1,w_2)=\sum_{w}p(w)\frac{\beta_{rel}}{\beta_{rel}+\mid I_{rel}(w,w_1)-I_{rel}(w,w_2)\mid} \tag{5}$$

其中参数 $1\geqslant\alpha_{rel}\geqslant 0$, $1\geqslant\beta_{rel}\geqslant 0$ 使用最大似然估计计算,分别由式(6,7)确定

$$\alpha_{rel}=\sum_{w_1}P(w_1)\sum_{w_2}P(w_2)\sum_{w}P(w)\mid I_{rel}(w_1,w)-I_{rel}(w_2,w)\mid) \tag{6}$$

$$\beta_{rel}=\sum_{w_1}P(w_1)\sum_{w_2}P(w_2)\sum_{w}P(w)\mid I_{rel}(w,w_1)-I_{rel}(w,w_2)\mid \tag{7}$$

定义 2 二元词组 w_1, w_2 之间的近似度定义为

$$\mathrm{sim}(w_1,w_2)=\sum_{rel}p(rel)\mathrm{sim}_{rel}(w_1,w_2) \tag{8}$$

基于词近似度,词类 C_1, C_2 之间的近似度定义为

$$\mathrm{sim}(C_1,C_2)=\frac{\sum\limits_{w_i\in C_1,w\in C_2}\mathrm{Count}(w_i)\mathrm{Count}(w_j)\mathrm{sim}(w_i,w_j)}{\sum\limits_{w\in C_1}\mathrm{Count}(w_i)\sum\limits_{w_j\in C_2}\mathrm{Count}(w_j)} \tag{9}$$

其中 $\mathrm{Count}(w_i)$, $\mathrm{Count}(w_j)$ 分别表示词 w_i 与 w_j 在语料中出现的数量。

1.2 词的聚类算法

词的聚类算法如下:(1)计算出任意两个词的语义近似度;(2)开始设置:词汇表里的任意一个词均假定为一个词类,总计 N 个词类(N 是词的总数目);(3)把语义近似度最大的两个词类合成为一个词类;(4)计算出其他词类和新合成的词类之间的语义近似度;(5)查验算法是否满足完结要求:词类的最大语义近似度小于事先确定的某个数值,或者词类合并个数满足算法的结束条件,如是,算法完结;否则,转(3)。

2 基于语义类的头驱动句法分析方法

头驱动的句子结构分析统计方法是典型的利用语义信息的句子结构分析方法。为了利用语义知识,句子语法生成式中的任何一个非终结符均引入词性/核心词等语义知识。然而语义知识的引进产生了稀疏数据难题。为了解决稀疏数据难题,该方法将语法规则的右边分解为三个主要组成:一个头成分、在右侧的几个短语组成和头左侧的几个短语组成,其中后面两个组成起修饰作用。即每个语法规则为

$$P(ht,hw)-L_m(lt_m,lw_m)\cdots L_1(lt_1,lw_1)H(ht,hw)$$
$$R_1(rt_1,rw_1)\cdots R_n(rt_n,rw_n) \tag{10}$$

式中:P 为非终结符号;H 为中心短语成分; L_1 为左边短语修饰成分; R_1 为右边短语修饰成分;hw, lw,

rw 都为短语成分的核心词；ht，lt，rt 相应表示核心词的词性。假定由非终结符号 P 生成中心短语 H，再分别以短语 H 为核心独立地生成所有左右两侧的短语(起修饰作用)。因而语法规则(10)的概率计算为

$$P_h(H \mid P(ht,hw)) \cdot \prod_{i=1}^{m+1} P_i(L_i(lt_i,lw_i) \mid H,P,h,\Delta_l(i-1)) \cdot \prod_{i=1}^{n+1} P_i(R_i(rt_i,rw_i) \mid H,P,h,\Delta_r(i-1)) \tag{11}$$

式中：L_{m+1} 和 R_{n+1} 分别表示左右两侧的相应中止符，$\Delta_l(i-1)$ 表示一种距离函数，用于对组成等信息的不足进行补偿。这里的距离函数主要补偿 3 类情形：(a)这个短语组成前面是否出现动词短语；(b)这个短语组成前面是否有短语组成；(c) 这个短语组成前面是否有标点符号。

使用词类语言模型(基于语义类)替换词的语言模型，可以缓解稀疏数据难题。令 $C(w)$ 代表 w 基于语义搭配关系的词聚类，则文法规则(10)就转换成如下形式，即有

$$\begin{aligned} &P(ht,C(hw)) - L_m(lt_m,C(lw_m)) \cdots L_1(lt_1,C(lw_1)) \\ &H(ht,C(hw)) \cdot R_1(rt_1,C(rw_1)) \cdots R_n(rt_n,C(rw_n)) \end{aligned} \tag{12}$$

而式(11)中的概率可近似为

$$\begin{aligned} &P_i(L_i(lt_i,lw_i) \mid H,P,h,\Delta_l(i-1)) \approx \\ &\quad P(lw_i \mid C(lw_i)) \cdot P_i(L_i(lt_i,C(lw_i)) \mid H,P,C(h),\Delta_l(i-1)) \end{aligned} \tag{13}$$

3 实验验证

3.1 词聚类实验

词聚类实验中采用的 Baseline 系统是一种较好的常规贪婪聚类方法[28]。本文采用《人民日报》中文标注语料库 1 月份语料和中文 PropBank2.0、中文 NomBank1.0 等中文语料库作为词聚类实验语料。《人民日报》中文标注语料库由富士通研究开发中心和北京大学计算语言学研究所共同加工《人民日报》1998 年中文语料制作。语言数据联盟公布了 CTB 中文树库，该树库是一个很好的中文句子结构分析测试和训练语料库。PropBank2.0 语料库是宾夕法尼亚大学在 TreeBank 5.1 中文句法结构分析语料库的基础上再标注了动词性谓词及其语义角色的中文语料库。而开发 NomBank1.0 中文语料库是为了弥补 PropBank 中文语料库只标注了动词性谓词的局限，它标注了 TreeBank 5.1 中文树库中的名词性谓词和其语义角色。《人民日报》中文 1 月份标注语料库共 120 万个词，现从其中选取约 90 万个词作为词的贪婪聚类算法训练用语料，其余约 30 万词作为贪婪聚类算法和基于语义相似度的聚类算法的开放测试语料，而中文 PropBank2.0、中文 NomBank1.0 等语料作为基于语义相似度聚类算法的训练语料。测试结果采用语言模型的困惑度作为评价指标，其定义为

$$PP_w = 2^{-\frac{1}{N_w}\sum_{i=1}^{N_w} \log_2 P(w_i \mid C(w_{i-2}),C(w_{i-1}))} \tag{14}$$

式中：困惑度 PP_w 为测试集概率分布几何平均的倒数；N_w 为测试语料中总词数；$C(w_{i-1})$ 代表词 w_{i-1} 所在的词类。一般来说，困惑度较小，语言模型更佳。

表 1 列出了两种聚类算法的聚类效果。从表 1 可以看出，基于语义相似度的词聚类算法的聚类效果明显好于常规贪婪聚类方法。

表 1 两种词聚类算法的聚类效果

Tab. 1 Clustering effects of two word clustering algorithms

聚类算法	困惑度
贪婪聚类算法	283
基于语义相似度的聚类算法	209.3

3.2 句法分析实验

句法分析试验数据取自中文 PropBank2.0 和中文 NomBank1.0。为了在训练语料、开发语料和测试语

料中平衡各种语料来源,参考 Xue[19] 的试验设置,分别利用汉语 PropBank2.0 和 NomBank1.0 中的各 40 个数据文件共 80 个数据文件当作句子结构分析试验的开发语料,各 648 个数据文件共 1296 个数据文件当作句子结构分析试验的训练语料。另利用 144 个数据文件当作句子结构分析试验的测试语料。在句子结构分析试验中,统计方法的主要参数均为利用极大似然法和平滑方法,从训练语料中计算出来。

采用句法分析召回率 R、句法分析准确率 P、句法分析的交叉括号 CB 和综合指标 F 值等 4 个典型的指标来评测句子结构分析试验的结果。评测指标的计算如下:精确率(P)表示句子结构分析结果中正确的短语结构在全部分析的短语结构中所占的比值;召回率(R)表示句子结构分析结果中正确短语结构在实际短语结构中所占的比值;综合指标:$F=(P\times R\times 2)/(P+R)$;交叉括号 CB 表示一个句子结构分析树与另外的句法树短语结构之间发生交界的平均短语结构数。

句子结构分析试验中取基于头驱动句子结构分析方法执行的 DBParser 作为基本方法。Petrov[29] 将自动发现隐藏的短语子块计算方法应用于汉语句子结构分析树库,基于正确的汉语分词,在 CTB5.0 汉语句子结构分析树库上获得了当时已知的基于正确汉语分词的汉语句子结构分析单语言模型的最高性能。表 2 列出了基本方法、Petrov 句子结构分析方法和基于语义类的头驱动句子结构分析统计方法的测试数据。

表 2 句子结构分析测试数据

Tab.2 Experimental data of syntactic parsing

方 法	精确率/%	召回率/%	F/%	交叉括号
Head-driven[11]	82.96	80.37	81.64	2.05
Petrov[29]	86.9	85.7	86.30	2.01
Improved model	88.73	88.26	88.49	1.85

稀疏数据难题严重影响句法结构分析模型的性能,本文采用基于语义聚类的平滑方法,比较成功地解决了稀疏数据问题,改进语言模型的各项性能比 Petrov 句法分析模型、著名的头驱动句法结构分析模型有较明显的提高。哈尔滨工业大学的曹海龙[1] 等提出了一个两级的中文句法分析方法,实验语料采用哈尔滨工业大学树库,实验结果为:召回率 88.0%,准确率 87.5%。这是已知中文句法分析的比较好的分析结果。与曹海龙等的句法分析方法相比较,模型 2 的性能也得到很大的提高。

本文还将改良的句子结构分析方法与头驱动的句子结构分析方法进行了组合,令分析树按照改良的句子结构分析方法、头驱动的句子结构分析方法分别计算的概率为 P_1,P_2,组合模型的概率 P 为

$$P=\lambda P_1+(1-\lambda)P_2 \tag{15}$$

其中 λ 取值为 $0\leqslant\lambda\leqslant 1$,通过改变 λ 的值,可以调整方法 1 和头驱动句法分析方法各自的贡献度。本文在开发集上进行实验,其综合指标 F 值如图 1 所示。当 λ 约为 0.6 时,组合句法结构分析方法的综合指标 F 的值最好。

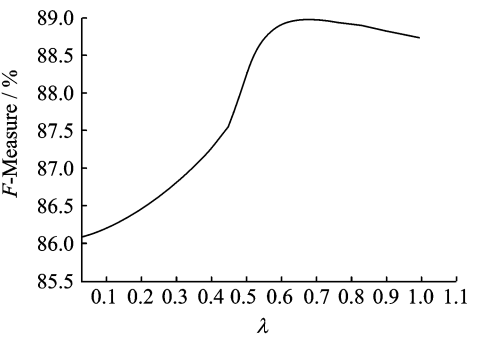

图 1 组合模型的综合指标 F 值

Fig.1 Comprehensive index F of combination model

4 结束语

句子结构分析是语义分析和短语结构分析的有机结合。句法分析不仅需做短语结构分析,比如句子主要短语组成分析、句子型式分析及短语成分联系分析等,并且还必需做相关的语义联系分析。对语义联系分析越深刻和全面,将更能够对短语结构上的种种语言问题给予合理和科

学的解答。在短语结构分析模型中融入诸如语义知识,词的语义搭配、词的语义依存和词的语义类等语义知识,将对语义和短语结构的分析和计算非常有用。为了缓解句子结构分析方法融入语义知识而引起的稀疏数据难题,本文在语义知识标注基础上提出了基于语义搭配关系的词聚类模型和算法,不但比较成功地缓解了稀疏数据难题,而且明显地提高了系统性能。

参考文献:

[1] 曹海龙. 基于词汇化统计模型的汉语句法分析研究[D]. 哈尔滨:哈尔滨工业大学,2006: 64-83.
Cao Hailong. Research on Chinese syntactic parsing based on lexicalized statistica model[D]. Harbin: Harbin University of Technology, 2006: 64-83.

[2] Vilares J, Alonso M A, Vilares M. Extraction of complex index terms in non-English IR: A shallow parsing based approach [J]. Information Processing and Management, 2008, 44(4):1517 - 1537.

[3] 刘水, 李生, 赵铁军,等. 头驱动句法分析中的直接插值平滑算法[J]. 软件学报, 2009, 20(11): 2915-2924.
Liu Shui, Li Sheng, Zhao Tiejun, et al. Directly smooth interpolation algorithm in head-driven parsing[J]. Journal of Software, 2009, 20(11):2915-2924.

[4] 代印唐,吴承荣,马胜祥,等. 层级分类概率句法分析[J]. 软件学报, 2011,22(2): 245-257.
Dai Yintang, Wu Chengrong, Ma Shengxiang, et al. Hierarchically classified probabilistic grammar parsing[J]. Journal of Software, 2011,22(2): 245-257.

[5] Aviran S, Siegel P H, Wolf J K. Optimal parsing trees for run-length coding of biased data[J]. IEEE Transaction on Information Theory, 2008, 54(2):841-849.

[6] Zhou Deyu, He Yulan. Discriminative training of the hidden vectors state model for semantic parsing[J]. IEEE Transaction on Knowledge and Data Engineering, 2009, 21(1): 66-77.

[7] 吴伟成,周俊生,曲维光. 基于统计学习模型的句法分析方法综述[J]. 中文信息学报, 2013, 27(3):9-19.
Wu Weicheng, Zhou Junsheng, Qu Weiguang. A survey of syntactic parsing based on statistical learning[J]. Journal of Chinese Information Processing, 2013, 27(3):9-19.

[8] 孙昂,江铭虎,贺一帆,等. 基于句法分析和答案分类的中文问答系统[J]. 电子学报, 2008,36(5): 833-839.
Sun Ang, Jiang Minghu, He Yifan, et al. Chinese question answering based on syntax analysis and answer classification[J]. Acta Electronica Sinica, 2008, 36(5): 833-839.

[9] 陈毅恒, 秦兵, 宋凡, 等. 基于 ontology 抽取优化初始选择的检索结果聚类[J]. 电子学报, 2008,36(12A):166-171.
Chen Yiheng, Qin Bing, Song Fan, et al. Search result clustering based on centroid optimization by ontology extraction[J]. Acta Electronica Sinica, 2008,36(12A):166-171.

[10] 袁里驰. 融合语言知识的统计句法分析[J]. 中南大学学报: 自然科学版, 2012, 43(3): 986-991.
Yuan Lichi. Statistical parsing with linguistic features[J]. Journal of Central South University: Natural Science, 2012, 43(3): 986-991.

[11] Collins M. Head-driven statistical models for natural language parsing[J]. Computational Linguistics, 2003, 29(4): 589-637.

[12] Jurafsky D, Martin J H. Speech and language processing[M]. New Jersey: Prentice Hall, 2009:210-265.

[13] Zhou M. A block-based dependency parser for unrestricted Chinese text[C] //Proceedings of the 2nd Chinese Language Processing Workshop. Hong Kong: Association for Computing Machinery, 2000: 78-84.

[14] Gao J F, Suzuki H. Unsupervised learning of dependency structure for language modeling[C]//Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics. Sapporo, Japan: Association for Computing Machinery, 2003: 521-528.

[15] Lai T B Y, Huang C N, Zhou M, et al. Span-based statistical dependency parsing of chinese[C] //Proceedings of the 6th Natural Language Processing Pacific Rim Symposium (NLPRS2001). Tokyo, Japan: IEEE Computer Society, 2001: 677-684.

[16] 李军辉. 中文句法语义分析及其联合学习机制研究[D]. 苏州:苏州大学, 2010: 64-103.
Li Junhui. Research on joint syntactic and semantic parsing for Chinese[D]. Suzhou: Soochow University, 2010: 64-103.

[17] 李军辉, 周国栋, 朱巧明, 等. 中文名词性谓词语义角色标注[J]. 软件学报, 2011, 22(8): 1725-1737.
Li Junhui, Zhou Guodong, Zhu Qiaoming, et al. Semantic role labeling in Chinese language for nominal predicates[J]. Journal of Software, 2011, 22(8): 1725-1737.

[18] 吴方磊,李军辉,朱巧明,等. 基于树核函数的中文语义角色分类研究[J]. 中文信息学报, 2011, 25(3): 51-58.

Wu Fanglei, Li Junhui, Zhu Qiaoming, et al. Tree kernel-based semantic role classification in Chinese language[J]. Journal of Chinese Information Processing,2011, 25(3): 51-58.

[19] Xue Nianwen. Labeling Chinese predicates with semantic roles[J]. Computational Linguistics, 2008, 34(2): 225-255.

[20] Bassiou N, Kotropoulos C. Long distance bigram models applied to word clustering[J]. Pattern Recognition, 2011, 44(1): 145-158.

[21] 宗慧, 刘金岭. 基于短文本信息流的热点话题检测[J]. 数据采集与处理, 2015, 30(2): 464-468.
Zong Hui, Liu Jinling. Hot topic detection based on short text information flow[J]. Journal of Data Acquisition and Processing, 2015, 30(2): 464-468.

[22] 宋文杰, 周俊生, 曲维光. 基于词典信息和网络百科的下位词获取[J]. 数据采集与处理, 2014, 29(5): 821-827.
Song Wenjie, Zhou Junsheng, Qu Weiguang. Chinese hyponymy extraction based on dictionary and encyclopedia resources [J]. Journal of Data Acquisition and Processing, 2014, 29(5): 821-827.

[23] Ido Dagan, Shaul Marcusb, Shaul Markovitchc. Context word similarity and estimation from sparse data[J]. Computer Speech and Language, 1995, 9(2): 123-152.

[24] 袁里驰. 基于相似度的词聚类算法和可变长语言模型[J]. 小型微型计算机系统, 2009, 30(5): 912-915.
Yuan Lichi. Word clustering based on similarity and vari-gram language model[J]. Journal of Chinese Computer Systems, 2009, 30(5): 912-915.

[25] Enhong Chen, Liu Shi, Dawei Hu. Probabilistic model for syntactic and semantic dependency parsing[C]// Proceedings of the 12th Conference on Computational Natural Language Learning. Manchester: Association for Computing Machinery, 2008:263-267.

[26] Surdeanu M, Johansson R,Meyers A, et al. The CoNLL-2008 shared task on joint parsing of syntactic and semantic dependencies[C]// Proceedings of the 12th Conference on Computational Natural Language Learning. Manchester: Association for Computing Machinery,2008:159-177.

[27] Duan Xiangyu, Zhao Jun,et al. Probabilistic models for action-based Chinese dependency parsing [C]// Proceedings of the 18th European Conference on Machine Learning. Warsaw, Poland: Springer, 2007: 559-566.

[28] Brown P F, Pietra V J D ,deSouza P V, et al. Class-based n-gram models of natural language[J]. Computational Linguistics,1992(18):467-479.

[29] Slav P,Klein D. Improved inference for unlexicalized parsing[C]// Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics. New York: Association for Computing Machinery, 2007:404-411.

作者简介:

袁里驰(1973-),男,博士,副教授,研究方向:自然语言处理,E-mail: yuanlichi@sohu.com。