

基于知网的汉语普通未登录词语义分析模型

张瑞霞¹ 杨国增² 闫新庆¹

¹(华北水利水电学院信息工程学院 河南 郑州 450011)

²(郑州师范学院数学系 河南 郑州 450044)

摘 要 提出一种基于知网的汉语普通未登录词语义分析模型,该模型以概念图为知识表示方法,以 2005 版知网为语义知识资源,首先参照知网知识词典对普通未登录词进行分词;然后综合利用知网中的知识词典等知识,通过词性序列匹配消歧法、概念图相容性判定消歧法、概念图相容度计算消歧法及语义相似度计算消歧法对中文信息结构进行消歧;最后根据所选择的中文信息结构生成未登录词的概念图,从而实现未登录词的语义分析。该模型在语义分析过程中一方面确定了未登录词中每个已登录词的词义,另一方面构造了该未登录词的语义信息,实验结果证明它可以作为普通未登录词语义分析的原型系统。

关键词 普通未登录词 知网 概念图 语义分析 语义相似度

中图分类号 TP391 文献标识码 A

HOWNET-BASED SEMANTIC ANALYSIS MODEL FOR GENERAL UNKNOWN CHINESE WORDS

Zhang Ruixia¹ Yang Guozeng² Yan Xinqing¹

¹(School of Information Engineering , North China University of Water Conservancy and Electric Power , Zhengzhou 450011 , Henan , China)

²(Department of Mathematics , Zhengzhou Teacher's College , Zhengzhou 450044 , Henan , China)

Abstract A HowNet-based semantic analysis model for general unknown Chinese words is put forward. It uses conceptual graphs as the method of knowledge representation and imposes HowNet version 2005 as the knowledge resource of the semantics. First, the model segments the general unknown Chinese words into some known words referring to the knowledge dictionary in HowNet; then by synthetically using the knowledge in HowNet such as the knowledge dictionary and so on, it disambiguates the Chinese message structures through the disambiguation methods of part of speech lists matching, conceptual graphs compatibility judgment, conceptual graphs compatibility computation and semantic similarity computation; Finally, it creates the conceptual graph of unknown Chinese words according to the selected Chinese message structure, and consequently completes the unknown Chinese words semantic analysis. In the model, the meaning of every known word in unknown words will be confirmed during semantic analysis on the one hand, and on the another hand the semantic information of the unknown words will be constructed. Experimental result proves that the model can be a prototype system for general unknown Chinese words semantic analysis.

Keywords General unknown words HowNet Conceptual graphs Semantic analysis Semantic similarity

0 引言

随着网络的广泛应用,涌现出了大量未登录词,关于未登录词的识别有很多研究,如文献[1-4]利用不同方法研究了未登录词的识别,文献[5-8]研究了未登录词人名的识别,文献[9,10]研究了地名的识别,文献[11,12]研究了专有名词的识别,由此可见目前关于未登录词的研究多集中在识别方面,特别是人名、地名、机构名等专有名词的识别,而关于普通未登录词语义分析的研究,如“毛衣店”、“检查费”、“给力”等语义分析的研究尚未见,因此设计合理的具有普遍性的未登录词语义分析模型有利于促进自然语言处理相关领域的发展。

鉴于上述原因,提出了基于知网的汉语普通未登录词语义分析模型,该模型以 2005 版知网为语义知识资源,以概念图为知识表示方法,选用《中文信息结构库》作为词语搭配规则库,

首先对中文信息结构进行形式化描述,然后根据知网词典对未登录词进行分词,接着在对中文信息结构进行消歧的基础上实现未登录词信息结构的选择,最后根据信息结构确定未登录词的语义。实验证明该模型是有效的。

1 知网和概念图

知网是一个以英汉双语所代表的概念以及概念的特征为基础的常识知识库,它主要描述了概念与概念之间以及概念所具有的特性之间的关系^[13]。根据它包含的知识的不同可以将其分为三部分,即:知识词典、义原特征文件、中文信息结构库;其

收稿日期:2012-03-29。河南省高校科技创新人才支持计划项目(2011HASTIT020)。张瑞霞,讲师,主研领域:自然语言处理,人工智能。杨国增,讲师。闫新庆,副教授。

中知识词典把每个词语以记录形式对其词性、概念项等进行了描述;义原特征文件一方面描述了同类义原间的层次关系,另一方面描述了不同义原间的语义关系;中文信息结构库描述了中文词语的各个组成部分之间由知网所规定的动态角色关系或属性,其揭示了中文的语言结构规律^[14];这些知识可作为未登录词语义分析的语义资源,即以知识词典作为基本词典,以中文信息结构库作为词语搭配规则库,参考众义原特征文件,通过不同方法对信息结构库中的规则进行消歧,从而为未登录词选择合适的信息结构,进而为未登录词中的每个已登录词选择合适的概念项,并最终确定未登录词的语义结构。

概念图是一种以语言学、心理学、哲学、逻辑学和数学为基础的知识表示方法^[15],它由概念结点和关系结点组成,概念结点表示实体、属性、状态和事件,关系结点表示概念间的关系,由于其建立在谓词逻辑上,因而能完全与自然语言相互翻译,表示出自然语言的语义^[16]。因为知网主要描述了概念与概念之间以及概念所具有的特性之间的关系,所以知网所包含的知识可以用概念图表示^[17],可以用概念图进行语义分析。不同于一般的概念图,基于知网的未登录词语义分析模型在用概念图作为知识表示方法时,概念图的结点只包含概念结点,把关系结点所表示的知识融入到了概念图的弧当中。

2 基于知网的普通未登录词语义分析模型

普通未登录词指的是未登录词中除去表示人名、地名、机构名等专有名词以外的其它未登录词。若没特别说明,下文中提到的未登录词均是指普通未登录词。

对未登录词进行语义分析,输入是未登录词,输出是采用概念图表示的未登录词的语义信息;基本思路是根据知网提供的知识,由已登录词的语义信息合成未登录词的语义信息,因此这里主要涉及了四个问题:

(1) 未登录词的分词问题,其目的是通过分词使未登录词由若干已登录词组成;

(2) 中文信息结构的表示问题,其目的是对中文信息结构进行形式化,使其能用概念图进行表示,以使用概念图进行语义分析;

(3) 消歧问题,包括中文信息结构的消歧和词义的消歧,其目的是确定已登录词的搭配关系和多义词的词义;

(4) 未登录词语义分析流程问题,其目的是合理安排语义分析过程,使未登录词能够得到有效的分析。

2.1 未登录词的分词

未登录词分词的目的是通过分词使未登录词由若干已登录词组成,因此分词方法需采用基于词典的分词方法,这里采用正向最大匹配和逆向最大匹配相结合的分词方法,若正向最大匹配和逆向最大匹配的分词效果一样,则保留两种分词结果。分词后的未登录词由若干已登录词组成,称这些已登录词为未登录词的小词,如未登录词“毛衣店”经过分词得到“毛衣店”,即“毛衣店”由小词“毛衣”和“店”组成。

2.2 中文信息结构的形式化

经过分词后的未登录词由若干小词组成,这些小词均可根据知网知识词典参照文献^[17]构造出其概念图,也称为词图;则对信息结构消歧,若能把信息结构中的语义结构也表示成概念图的形式,则可应用图论的相关理论进行计算。因此首先对

中文信息结构进行形式化,主要表现在四方面:

(1) 在信息结构中,对概念的描述采用知网知识词典中概念项的描述形式,这样信息结构即可采用概念图的形式表示;

(2) 用 L 或 R 代替信息结构描述中的左箭头或右箭头,来表示词语间的修饰关系;

(3) 按照信息结构的语义结构描述式中是否有具体词语及主义原是否是义原树的叶结点,把信息结构分为一、二、三优先级,按照优先级别依次对中文信息结构进行消歧。

(4) 对信息结构中的词例进行词性、词义、角色标注。

以下代码给出了一信息结构的示例,其中 ID 表示与之对应的标识号,SYN_S 表示句法结构,SEM_S 表示语义结构,Example 列举了符合该信息结构的部分词例,CLASSPRO 表示其优先级。

ID = 38

SYN_S = V L N

SEM_S = { ((物质 [possession] L < 买 > } L [location1] (组织/场所)

Example = 百货/N/00200 - [location1] - 商店/N/073032,书/N/078158 - [location1] - 店/N/022012,鞋/N/092866 - [location1] - 店/N/022012,服装/N/028601 - [location1] - 店/N/022012

CLASSPRO = 2

2.3 中文信息结构的消歧

经过形式化描述的信息结构可用概念图表述出 SEM_S 部分,实际上对信息结构消歧即是对 SYN_S 和 SEM_S 的消歧,因此信息结构的消歧可根据词性序列及概念图表示的语义信息进行消歧,这里提出了四种消歧方法:即词性序列匹配消歧法、概念图相容性判定消歧法、概念图相容度计算消歧法及语义相似度计算消歧法。这四种消歧方法按照由简单到复杂、由定性分析到定量计算的指导思想进行设计,接下来对四种消歧方法进行详细探讨。

2.3.1 词性序列匹配消歧

即若信息结构的词性序列 SYN_S 与未登录词小词的词性序列完全匹配,则该信息结构可能成为未登录词的信息结构;否则,该信息结构不能成为未登录词的信息结构。如“毛衣店”对应的词性序列为“N N”,经过词性序列匹配消歧后保留了 74 个信息结构,与信息结构总数 270 相比数量大大减小。

2.3.2 概念图相容性判定消歧

在介绍概念图相容性判定消歧法之前,先介绍基本定义。设 G_1 与 G_2 均为概念图,其中 $G_1 = \langle V_1, E_1 \rangle$, $G_2 = \langle V_2, E_2 \rangle$, $\forall v \in (V_1 \cup V_2)$, $s(v)$ 表示结点 v 所代表的义原:

定义 1 G_1 等相容于 G_2 : 若存在双射函数 $f: V_1 \rightarrow V_2$, 对于 $v_i, v_j \in V_1$, $\langle v_i, v_j \rangle \in E_1$ 当且仅当 $\langle f(v_i), f(v_j) \rangle \in E_2$, 并且 $s(v_i) = s(f(v_i))$, $s(v_j) = s(f(v_j))$, 则称 G_1 等相容于 G_2 , v_i 等相容于 $f(v_i)$, v_j 等相容于 $f(v_j)$, 分别记作 $G_1 \approx G_2$, $v_i \approx f(v_i)$, $v_j \approx f(v_j)$ 。

从图等相容的定义易知: 图的等相容性具有自反性、对称性和传递性。

定义 2 G_1 上相容于 G_2 : 若存在双射函数 $f: V_1 \rightarrow V_2$, 对于 $v_i, v_j \in V_1$, $\langle v_i, v_j \rangle \in E_1$ 当且仅当 $\langle f(v_i), f(v_j) \rangle \in E_2$, 并且 $s(v_i)$, $s(v_j)$ 分别是 $s(f(v_i))$, $s(f(v_j))$ 的子孙义原, 则称 G_1 上相容于 G_2 , v_i 上相容于 $f(v_i)$, v_j 上相容于 $f(v_j)$, 分别记作 $G_1 \leq G_2$, $v_i \leq f(v_i)$, $v_j \leq f(v_j)$ 。

定义 3 G_1 局部等相容于 G_2 : 若 G_1 存在子图 G_1' 等相容于 G_2 , 则称 G_1 局部等相容于 G_2 , 记作 $G_1 \approx G_2$ 。

定义4 G_1 局部上相容于 G_2 : 若 G_1 存在子图 G_1' 上相容于 G_2 , 则称 G_1 局部上相容于 G_2 , 记作 $G_1 \mid \leq G_2$ 。

由以上四个定义可知, 图的上相容性具有自反性、反对称性和传递性, 是一偏序关系, 同理图的局部等相容关系、局部上相容关系均是偏序关系; 同时若 $G_1 \approx G_2$, 则有 $G_1 \mid \approx G_2$; 若 $G_1 \leq G_2$, 则有 $G_1 \mid \leq G_2$ 。

概念图相容性判定消歧主要是通过图相容性的判定进行消歧, 即若未登录词中的词图与 SEM_S 中对应的词图满足一定的相容关系, 则该信息结构有可能成为未登录词的信息结构, 否则不能成为其信息结构; 根据图的相容关系, 图相容性判定消歧法可分为图局部等相容判定消歧法、图局部上相容判定消歧法。

如“毛衣”根据知网词典构造“毛衣”和“店”的概念图分别如图1、图2所示, 再经过概念图相容性判定消歧后, 74 个信息结构只剩下 10 个, 如表1所示。



图1 “毛衣”的概念图

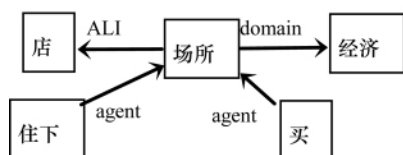


图2 “店”的概念图

表1 通过概念图相容性判定消歧的信息结构表

ID	SEM_S
1	(万物 [compound] L (万物))
7	(万物 [and] L (万物/部件))
29	{{ (物质 [patient] L <处理>) L [agent1] } (人/组织)}
38	{{ (物质 [possession] L <买>) L [location1] } (组织/场所)}
39	{{ (物质/事情/事务 [patient] L <处理>) L [agent1] } (组织/场所)}
40	{{ (物质/事情/事务 [instrument] L <消闲>) L [location1] } (组织/场所)}
41	{{ (物质 [content] L <医治>) L [location1] } (组织/场所)}
50	(万物 [possession] L (万物))
66	{{ (万物 [relevant] L <导致>) [cause] L (万物/时间)}
94	(物质 [contrast] L (万物/部件))

2.2.3 概念图相容度计算消歧法

表1中的信息结构与“毛衣”“店”均满足相容性, 如何衡量相容性的强弱? 因此引入概念图相容度的概念来量化概念图的相容性, 即相容度大的概念图相容性强。设概念图 G_1, G_2 令其相容度为 $com(G_1, G_2)$, 则当 G_1, G_2 不具有相容关系时令 $com(G_1, G_2) = 0$; 当 $G_1 \approx G_2$ 时, 令 $com(G_1, G_2) = 1$; 接下来讨论当 $G_1 \mid \leq G_2$ 时 $com(G_1, G_2)$ 的计算方法。

设 G_1, G_2 是词图, 其中 G_1' 是 G_1 的子图, 函数 f 是 G_1' 到 G_2 同构映射:

当 G_1' 中不含有 G_1 的中心义原结点时, 说明 G_1' 是辅助说明 G_1 的, 而含有 G_1 中心义原结点的子图与 G_2 不具有相容关系, 即代表 G_1 与 G_2 不具有相容关系, 因此令 $com(G_1, G_2) = 0$ 。

当 G_1' 中含有 G_1 的中心义原结点时, 即代表 G_1 与 G_2 具有相容关系, 设 G_1 的中心义原结点为 v_{1i} , 则 G_1 和 G_2 的相容度 $com(G_1, G_2)$ 由具有相容关系结点的相容度加权之和构成, 如式

(1) 所示:

$$com(G_1, G_2) = comVex(v_{1i}, f(v_{1i})) + \sum_{i=2}^n (comVex(v_{1i}, f(v_{1i})) \times \frac{shortLength + \eta_1}{gDia + \eta_1}) \quad (1)$$

在式(1)中 $n = |V_2|$; $shortLength$ 是在 G_1 中 v_{1i} 到 v_{11} 的最短路径长度; $gDia$ 表示 G_1 的直径; η_1 是调节参数; $comVex(v_{1i}, f(v_{1i}))$ 表示结点 v_{1i} 与结点 $f(v_{1i})$ 的相容度, 因为结点距离中心结点越远其相容度对整体相容度影响越小, 所以对 $comVex(v_{1i}, f(v_{1i}))$ 使用 $\left(\frac{shortLength + \eta_1}{gDia + \eta_1}\right)$ 进行修正。

计算 $comVex(v_{1i}, f(v_{1i}))$ 时, 当两个结点表示词语时, 若表示相同词语, 则 $comVex(v_{1i}, f(v_{1i})) = 1$, 否则 $comVex(v_{1i}, f(v_{1i})) = 0$; 当两个结点表示义原时, 若表示相同义原, 则 $comVex(v_{1i}, f(v_{1i})) = 1$; 若表示的义原不具有子孙关系, 则它们是不相容的, 此时 $comVex(v_{1i}, f(v_{1i})) = 0$; 若表示的义原具有子孙关系, 那么两个结点的相容度与其相似度有一定关系, 即两义原越相似, 其相容度越大, 此时参照相似度计算方法计算 $comVex(v_{1i}, f(v_{1i}))$, 如式(2)所示:

$$comVex(v_{1i}, f(v_{1i})) = \frac{\min(\text{depth}(s(v_{1i})), \text{depth}(s(f(v_{1i}))))}{\text{depthTree}} \times \frac{\lambda_1}{\lambda_1 + \text{dis}(s(v_{1i}), s(f(v_{1i})))} \quad (2)$$

在式(2)中, $s(v_{1i})$ 表示结点 v_{1i} 所代表的义原; $\text{depth}(s(v_{1i}))$ 表示义原 $s(v_{1i})$ 在义原树上的深度; \min 取两者的最小值; depthTree 表示 $s(v_{1i})$ 所在义原树的高度; λ_1 为调节参数; $\text{dis}(s(v_{1i}), s(f(v_{1i})))$ 计算两义原的距离即 $s(v_{1i})$ 在义原树上到 $s(f(v_{1i}))$ 的最短路径长度。

以上讨论了, 当 $G_1 \mid \leq G_2$ 时 $com(G_1, G_2)$ 的计算方法; 同样, 当 $G_1 \mid \leq G_2$ 时 $com(G_2, G_1)$ 为何值? 按照相容性的定义, 因为 G_2, G_1 具有相容关系并且与 G_1, G_2 所具有的相容关系一致, 所以有 $com(G_2, G_1) = com(G_1, G_2)$, 为了表示出相容关系的方向性, 故令 $com(G_2, G_1) = -com(G_1, G_2)$ 。

因为两个图的相容性越强其相容度越大, 而两个图的相容度越大, 则其语义结构越相似, 所以概念图相容度计算消歧法是通过计算未登录词的概念图与 SEM_S 概念图的相容度进而选择相容度最大的信息结构成为最优信息结构。因此该方法关键是计算未登录词和 SEM_S 的相容度, 即计算两者图结构的相容度。设未登录词中含有 m 个词语, 各个词图组成的集合 $GS_1 = \{G_{1i} \mid 1 \leq i \leq m\}$, SEM_S 各个概念图组成的集合 $GS_2 = \{G_{2j} \mid 1 \leq j \leq n\}$, 则当且仅当 $n = m$ 时, SEM_S 可能成为未登录词的信息结构, 且若 G_{2i} 为 SEM_S 中心词语的概念图, 则有 G_{1i} 为未登录词中心词语的概念图, 中心词语概念图的相容度对整体相容度影响要高于非中心词语概念图对整体相容度的影响; 而关于非中心语对整体相容度影响与其距中心语的距离有关, 即距离中心语越近, 影响越大; GS_1 和 GS_2 的相容度 $comGS(GS_1, GS_2)$ 如式(3)所示:

$$comGS(GS_1, GS_2) = com(G_{1i}, G_{2i}) + \sum_{k=1}^{i-1} (com(G_{1k}, G_{2k}) \times \frac{\lambda_2}{\lambda_2 + \text{disG}(G_{2k}, G_{2i})}) + \sum_{k=i+1}^m (com(G_{1k}, G_{2k}) \times \frac{\lambda_2}{\lambda_2 + \text{disG}(G_{2k}, G_{2i})}) \quad (3)$$

在式(3)中 $com(G_{1i}, G_{2i})$ 表示概念图 G_{1i} 和 G_{2i} 的相容度, 按照式(1)计算; λ_2 为调节参数; $\text{disG}(G_{2k}, G_{2i})$ 表示 G_{2k} 与 G_{2i} 的语义距

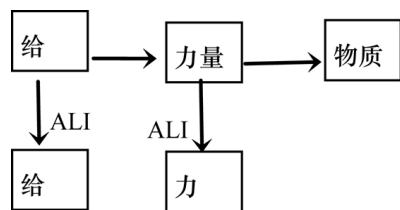


图5 “给力”的概念图

3.2 实验结果分析

本实验的实验集由两部分组成,第一部分来自《PFR 人民日报标注语料》,从中统计出 8321 个普通未登录词;第二部分来自自然语言处理开发平台热门下载中《基于网页的词频统计》中所下载的网页中的关键词,并用中科院研究所的分词系统 IC-TCLAS 对其进行分词,然后统计出 13957 个普通未登录词。

为了对进行实验结果评价,提出准确率,其计算方法如下:

$$\text{准确率} = \frac{\text{语义分析正确的未登录词的总数}}{\text{实验集中未登录词的总数}} \times 100\%$$

由于未登录词经分析后得到的是关于其语义的概念图,而对语义概念图正确性的评价目前尚无自动评价系统,因此一方面本实验采用人工方式进行判定,即如果未登录词语义概念图所表达的语义与人对未登录词语义理解一致,则认为该未登录词得到正确的语义分析,否则认为其语义分析错误。经过人工判定,在 22278 个未登录词中有 17666 个语义分析正确,准确率为 79.3%。影响准确率的主要因素有以下三点:

(1) 关于含有动词的未登录词,由于动词的灵活性很高,分析的准确率受影响,如“砍价”。

(2) 关于缩略语式的未登录词,由于缺乏缩略语与词典词语的对应关系,所以分析的准确率很低,如未登录词“党代会”。

(3) 关于成语式的未登录词,由于《中文信息结构库》的局限性,分析的准确率不高,如“大器晚成”。

另一方面本实验将经过语义分析的未登录词用于未登录词语义相似度计算取得了较好效果^[19]。

4 结 语

以 2005 版知网为语义资源提出了普通未登录词语义分析模型。该模型首先参照知网知识词典对普通未登录词进行分词,然后利用知网中的知识词典、中文信息结构库、义原特征文件为语义资源,以概念图为知识表示方法,采用词性序列匹配消歧法、概念图相容性判定消歧法、概念图相容度计算消歧法及语义相似度计算消歧法对中文信息结构进行消歧,最后根据中文信息结构生成未登录词的概念图,从而实现了未登录词的语义分析。

目前,基于知网的普通未登录词语义分析仍处在模型阶段,在下一步的工作中,可以参考现有研究成果加入专有名词语义分析模块;为了提高含有动词未登录词的分析效率,可以加入基于统计分析的动词句法结构分析模块;关于缩略语式的未登录词语义分析,可以加入常用词与其缩略语对应表,设计缩略语式未登录词语义分析模块;关于概念图的生成,目前仅仅是根据信息结构合成各个词图,还应该添加概念图的修剪模块等等。

参 考 文 献

[1] 吕雅娟,赵铁军,杨沐昀. 基于分解与动态规划策略的汉语未登录词识别[J]. 中文信息学报, 2001, 15(1): 28-33.

[2] 周蕾,朱巧明. 基于统计和规则的未登录词识别方法研究[J]. 计算机工程, 2007, 33(8): 196, 198.

[3] 王蕾,杨季文. 汉语未登录词识别现状及一种新识别方法介绍[J]. 计算机应用与软件, 2007, 24(8): 213-215.

[4] 都菁,熊海灵. 基于论坛语料识别中文未登录词的方法[J]. 计算机工程与设计, 2010, 31(3): 630-633.

[5] 张华平,刘群. 基于角色标注的中国人姓名自动识别研究[J]. 计算机学报, 2004, 27(1): 85-91.

[6] 张锋,樊孝忠,许云. 基于统计的中文姓名识别方法研究[J]. 计算机工程与应用, 2004, 40(53): 54-57.

[7] 孟伟涛,张蕾,张晓李. 一种基于位置概率模型的中文人名识别方法[J]. 计算机应用与软件, 2008, 25(4): 186-189.

[8] 戴播,毛奇,袁春风. 一种基于共坐标上升算法的人名识别方法[J]. 计算机应用与软件, 2010, 27(4): 7-9.

[9] 李诺,张全. 利用地名用字分析的中文地名识别处理[J]. 计算机工程与应用, 2009, 45(28): 230-232.

[10] 鞠久朋,张伟伟,宁建军. CRF 与规则相结合的地理空间命名实体识别[J]. 计算机工程, 2011, 37(7): 220-222, 215.

[11] 冯元勇,孙乐,李文波. 基于单字提示特征的中文命名实体识别快速算法[J]. 中文信息学报, 2008, 22(1): 104-110.

[12] 胡学营,刘慧,陆汝占. 搜索引擎用户查询中的复杂专有名词识别[J]. 计算机工程与应用, 2008, 44(19): 153-155.

[13] 董振东,董强. 知网——知网简介[R/OL]. <http://www.keenage.com>.

[14] 董振东,董强. 关于知网——中文信息结构库[R/OL]. <http://www.keenage.com>.

[15] Sowa J F. Conceptual Graphs[R/OL]. <http://www.jfsowa.com/cg/index.htm>.

[16] 张蕾,李学良. 概念结构及其应用[D]. 西安: 西北工业大学, 2001.

[17] 张瑞霞,肖汉. 基于知网的词图构造[J]. 华北水利水电学院学报, 2008, 29(3): 53-56.

[18] 张瑞霞,朱贵良,杨国增. 基于知识图的汉语词汇语义相似度计算[J]. 中文信息学报, 2009, 23(3): 116-120.

[19] 张瑞霞,杨国增,吴慧欣. 基于《知网》的汉语未登录词语义相似度计算[J]. 中文信息学报, 2012, 26(1): 16-21.

(上接第 51 页)

[2] Aleksovski Z, Klein M, Kate W, et al. Matching unstructured vocabularies using a background ontology[C]//Knowledge Engineering and Knowledge Management(EKAW) 2006: 182-197.

[3] Madhavan J, Bernstein P A, Rahm E. Generic schema matching with Cupid[C]//27th Intl. Conference on Very Large Databases (VLDB), Rome, Italy, Sep. 2001: 49-58.

[4] Gruber T R. A Translation Approach to Portable Ontology Specifications[J]. Knowledge Acquisition, 1993, 5(2): 199-220.

[5] Borst W N. Construction of Engineering Ontologies for Knowledge Sharing and Reuse[D]. University of Twente, Enschede, 1997.

[6] Jaya A, Uma G V. An intelligent system for semi-automatic story generation for kids using ontology[C]//Proceedings of the Third Annual ACM Bangalore Conference, 2010: 1-6.

[7] Caracciolo C, Euzenat J, Hollink L, et al. Results of the Ontology Alignment Evaluation Initiative 2008[C]//3rd ISWC workshop on ontology matching (OM), Karlsruhe (DE), 2008: 73-119.

[8] SourceForge.net. Jazzy - Java Spell Check API[CP/OL].

[9] 唐杰,梁邦勇,李涓子. 语义 Web 中的本体自动映射[J]. 计算机学报, 2009, 29(11): 1956-1976.