第16卷 第5期
2006年5月

计算机技术与发展
COMPUTER TECHNOLOGY AND DEVELOPMENT

Vol.16 No.5
May 2006

基于义原同现频率的汉语词义排歧系统

刘亚清,张 瑾,于纯妍
(大连海事大学 计算机科学与技术学院,辽宁 大连 116026)

摘 要:词义排歧在自然语言处理领域占有重要地位。词义排歧的精确率依赖于排歧知识的完备性。但是目前使用的基于词典的和基于语料库的词义排歧方法来获取排歧知识的效果都不令人满意。文中借助《知网》,以义原同现频率矩阵作为排歧知识,在其基础上设计并实现了一个基于义原同现频率的汉语词义排歧系统,大大地提高词义排歧的精确率。

关键词:自然语言处理;词义排歧;义原
中图分类号:TP391.12 **文献标识码**:A **文章编号**:1673-629X(2006)05-0184-02

A Chinese Word Sense Disambiguation System Based on Primitive CO-Occurrence Data

LIU Ya-qing, ZHANG Jin, YU Chun-yan
(Institute of Computer Science and Technology, Dalian Maritime University, Dalian 116026, China)

Abstract: Word sense disambiguation has always been important in natural language processing. The precision rate of word sense disambiguation depends on the completeness of disambiguation knowledge. Most methods currently which are based on dictionary and on corpora are not satisfiable. This paper introduces an automatic system of Chinese word sense disambiguation which is based on HOWNET and takes primitive co-occurrence data matrix as the disambiguation knowledge. The result of experimentation proves that the the precision rate of word sense disambiguation is higher.

Key words: natural language processing; word sense disambiguation; primitive

词义排歧,就是对一个多义词,根据它的上下文给出它在这个上下文环境所对应的惟一的语义编码。词义排歧研究的目的在于使计算机能够在特定的语言环境下判断出多义词的词义。词义排歧属于自然语言处理领域,它在自然语言处理领域中扮演着承上启下的角色,是机器翻译、信息检索、主题内容分析和文本处理等课题研究的前提和基础。

1 词义排歧模型

词义排歧是自然语言处理领域中的一个中间环节,它是在自动分词、词性标注的基础上在特定的语言环境下对多义词的词义进行自动标注的过程。由图1可知,词义排歧的过程就是利用词义排歧系统,根据词义排歧知识对待标注文本(该文本已经过自动分词、词性标注)进行词义排歧的过程。

2 词义排歧的知识

词义排歧的知识是词义排歧系统对多义词进行排歧时所凭借的理论知识。笔者所采用的排歧知识是义原同现频率矩阵[1],建立过程如下所述。

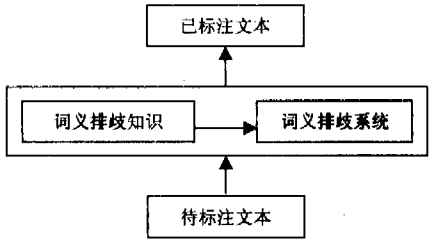

图1 词义排歧模型

首先作出如下约定:矩阵中的行和列都为《知网》中的基本义原,其值为两个义原之间的同现频率。其中,DC_{ij}表示DC_i与DC_j两个义原之间的同现频率,$DC_{ij} = DC_{ji}$;在训练语料中从未在同一句子中出现的两个义原同现频率记为0。

第一步:初始化义原同现频率矩阵DC,将其中每一单元均置零;

第二步:以句子为单位划分训练语料;

第三步:对于每一个句子S中的每一个词语,标注其义原扩展集γE;

第四步:对S中任意两个词的义原扩展集γE_i,γE_j中的任意两个义原集γS_{im},γS_{jn}中的任意两个义原DC_{imp},

收稿日期:2005-08-24
作者简介:刘亚清(1979-),男,山西大同人,硕士,讲师,研究领域为自然语言处理。

DC_{jmq} 计算 $W(DC_{imp}) \times W(DC_{jmq})$，其中 $W(D_{xyz}) = \frac{1}{|\gamma E_x| \times |\gamma S_{xy}|}$，$|\gamma E_x|$ 表示词 x 的义原集的个数，而 $|\gamma S_{xy}|$ 表示词 x 义原扩展集中 y 义原集中义原的个数。

第五步:将 $W(DC_{imp}) \times W(DC_{jmq})$ 累加到 DC 中相对应的单元上，就形成义原同现频率矩阵。

3　词义排歧系统开发的思路

首先对待标注的语料按句子为单位进行划分:

(1)首先根据《多义词词林》[2]生成多义词词典;

(2)对于句子中的每一个词,扫描多义词词典,如果发现命中,则将该词标注为待排歧词,其余的词语为特征词,所有特征词的集合组成特征词表;

(3)利用《知网》[3]生成义原释义表;

(4)利用义原释义表把特征词和待排歧词分别用相应的释义义原来表示,并根据义原在各自概念释义中的类型划分为第一独立义原、其他独立义原、关系义原、符号义原 4 类[3]。然后利用义原同现频率矩阵分别对每类义原计算特征词与待排歧词的每一个词义的相关系数,将相关系数最大的词义确定为该待排歧词在当前语言环境中的词义。然后对下一条句子进行相同的处理,如果最后一条句子处理完成,则排歧过程结束。

基于词义排歧系统的总体开发思路,笔者将系统划分为 3 个模块:确定待排歧词模块、词语 - 义原转换模块、确定多义词义项模块。下面分别对这 3 个模块进行介绍。

3.1　确定待排歧词模块

本模块主要解决的问题是如何从给定的句子中确定待排歧词。本模块实现的思路如下:

(1)首先对给定的句子进行分词,提取第一个词;

(2)在多义词词典中查找是否有命中的词语,如果命中,将该词标注为待排歧词;

(3)判断句子是否结束,如果结束,则转到步骤(4);否则提取下一个词语,转到步骤(2);

(4)如果该句子中的词没有一个命中,说明句子中无多义词,提取下一条句子,转到步骤(1)。

3.2　词语 - 义原转换模块

本模块主要解决的问题是如何将特征词表中的词以及待排歧词利用《知网》转换为相应的义原表示。需要特别指出的是,在具体得到每个概念的义原时,笔者主要利用了《知网》中的 HowNetSystem 软件(如图 2 所示),然后生成义原释义表。

本模块实现的思路如下:

(1)提取第一个特征词,转到步骤(2);

(2)在概念释义表中查找该特征词,将命中的该词的所有义原信息保存下来;

(3)提取下一个特征词,如果为空,则转到步骤(4);否则转到步骤(2);

(4)特征词词语 - 义原转换完毕。

待排歧词的词语 - 义原转换方法同特征词类似,不同之处在于因待排歧词在概念释义表中不只一个义项,故待排歧词有多少个义项,相应就会得到多少个义原表示。

3.3　确定多义词义项模块

本模块在充分利用上述模块结果的基础上,最终确定多义词在特定环境下的义项。其实现思路如下:

(1)提取待排歧词的第一个义项以及相应的释义义原,转到步骤(2);

(2)提取所有特征词的的释义义原,转到步骤(3);

(3)分别计算待排歧词当前义项同所有特征词的第一独立义原、其他独立义原、关系义原、符号义原之间的相关系数[4];

(4)赋予第一独立义原、其他独立义原、关系义原、符号义原的权重分别为:0.5,0.2,0.17,0.13[5],然后求和得到待排歧词当前义项同特征词的相关系数;

(5)查找待排歧词下一个义项,如果为空,则转到步骤(6),否则提取该义项以及相应的释义义原,转到步骤(3);

(6)比较所得到的相关系数并将相关系数最大者所表征的词义标注给待排歧词,排歧过程结束。

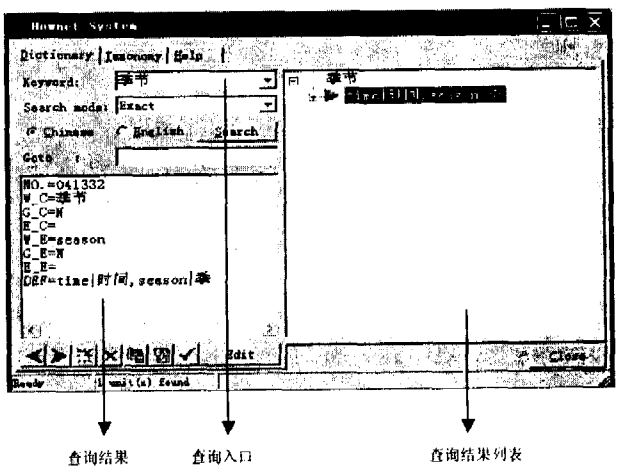

图 2　HowNetSystem 软件使用界面

4　实　验

本系统选用 C 语言作为程序开发语言,选用 Windows2000 作为操作系统平台。由于本系统研究的对象是汉语词语,因此主要涉及的是字符串,而 C 语言对于字符的处理非常灵活,有大量的字符串处理函数供调用。另外,作为实验系统,C 语言中的结构体数组能够非常直观清晰地表示一些词典资源,既方便了程序设计又增加了程序的可读性。

根据笔者开发的词义排歧系统对“材料”、“路线”等 20 多个多义词进行词义排歧,取得令人欣慰的排歧结果。在进行测试时,分别从训练语料中和训练语料之外的同类型的新闻语料中选取了不同歧义词的实例 100 个,并利用如下公式计算词义排歧正确率。正确率 = 正确标注义项号的样本个数/测试集中样本总数。测试结果为:开放测

(下转第 188 页)

第五步,将力产生器插入到适当的单元。

最后,渲染粒子。

2.3 蜡烛火焰的编程实现

有了粒子的运动模型,利用程序的迭代循环就可以刷新每个粒子在屏幕上的显示,形成帧动画。但在编程时还应考虑怎样提高粒子的真实感和实时性,文中通过以下方法来实现:

(1)建立粒子类,将粒子的属性和刷新方法集成到类中。

(2)在主程序中建立粒子类的实例数组,在刷新环境属性的同时,刷新每个粒子的属性。

(3)用标志位标识粒子。高级的粒子系统表达的物体越真实,粒子的数量就越多,同时带来的是大量的时间和空间消耗。因此设计粒子系统的数据结构非常重要,否则将影响粒子系统的刷新率。

实际应用中,大多数的性能问题都是由内存管理的失败引起的,因此,应尽量少地执行分配和释放内存的内存操作。当粒子系统中的某一个粒子死亡时,不必将它从内存中释放。而是用一个标志位来记录它是死亡还是重新初始化,当所有的粒子都标记为死亡或整个粒子系统完成任务后,才将所有粒子占用的存储空间同时释放。

(4)利用 OpenGL 显示列表优化图形的绘制。OpenGL 中为加快图形的绘制速度提供了一套显示列表的接口。显示列表是一组存储起来用于稍后执行的 OpenGL 命令。激活一个显示列表后,就按照显示列表中预先排好的次序执行其存储的命令,在程序中可以自由地使用这些命令。与子程序不同的是,这些命令是经过编译的,执行效率高,从而可以有效地提高 OpenGL 的绘图性能。按照粒子系统绘制流程,状态刷新后的粒子将被逐一地全部映射到屏幕上,逐一对粒子进行计算和映射是十分耗时的过程,因此使用显示列表技术处理粒子维护和控制模块中的模拟粒子老化和消隐过程,可以大大减少绘制时间,并使粒子消隐绘制的时间复杂度只与绘制区域相关,而与粒子数量无关。显示列表实际上是一系列命令的高速缓存,而不是在内存中的动态数据库,故不必进行内存管理,也不占用内存资源,能大大提高绘制性能。

(5)碰撞检测,算法如下:

判断粒子的当前位置是否位于反弹面(如地面)之后,若粒子落到反弹面之后,则标识该粒子消亡或予以新生(反弹此粒子)。

3 结 论

文中用粒子系统模拟蜡烛的燃烧,燃烧过程连续,具有一定的真实感。动态过程力域的应用,保证了蜡烛火焰的真实感,用标志位来标识粒子的死亡和用 OpenGL 显示列表技术提高绘制效率。实验结果表明,该方法满足实时性和真实感的要求。

参考文献:

[1] Reeves W T. Particle Systems - A technique for modeling a class of fuzzy objects[J]. Computer Graphics, 1983, 17(3): 359 - 376.

[2] Karl S. Particle Animation and Rendering Using Data Parrel Computation[J]. Computer Graphics, 1990, 24(4): 405 - 413.

[3] Scholl A, Szeliski R, Salesin D H, et al. Video Texture[A]. Proceedings of ACM SIGGRAPH 2000 Conference[C]. New Orleans, LA: ACM Press, 2000. 489 - 498.

[4] Fedkiw R, Stam J, Jensen H. Visual Simulation of Smoke [A]. Proceedings of ACM SIGGRAPH 2001 Conference[C]. New York, NY, USA: ACM Press / ACM SIGGRAPH, 2001. 15 - 22.

[5] van Wijk J J. Image Based Flow Visualization[A]. Proceedingss of the 29th annual conference on Computer Graphics and interactive techniques[C]. New York, NY, USA: ACM Press, 2002. 745 - 754.

[6] 詹云开,罗世彬,贺汉根.用粒子系统理论模拟虚拟场景中的火焰和爆炸过程[J]. 计算机工程与应用, 2001(5): 91 - 92.

[7] 张 芹,吴慧中,谢隽毅,等.基于粒子系统的火焰模型及其生成方法研究[J]. 计算机辅助设计与图形学学报, 2001, 13(1): 78 - 83.

[8] Nguyen D, Fedkiw R, Jensen H. Physically Based Modeling and Animation of Fire[A]. Proceedings of the 29th International Conference on Computer Graphics and Interactive Techniques[C]. New York, NY, USA: ACM SIGGRAPH, 2002. 721 - 728.

[9] Lamorlette A, Foster N. Structural Modeling of Flames for a Production Environment[A]. Proceedings of the 29th Annual Conference on Computer Graphics and Interactive Techniques [C]. New York, NY, USA: ACM SIGGRAPH, 2002. 729 - 735.

(上接第 185 页)

试正确率是 79%,封闭测试的正确率为 86%。相比较传统的基于义原同现频率的汉语词义排歧方法,开放测试的正确率提高了 8%,封闭测试的正确率提高了 11%。

参考文献:

[1] 杨尔弘,张国清,张永奎.基于义原同现频率的汉语词义排歧方法[J].计算机研究与发展,2001(7):833 - 838.

[2] 梅加驹.同义词词林[M].上海:上海辞书出版社,1983.

[3] 董振东.知网[EB/OL]. http://www.keenage.com,2000.

[4] 刘亚清.基于词义的汉语排歧方法研究[D].南京:南京理工大学,2002.

[5] 程 莉,卢正鼎,文坤梅,等.基于语义的模糊匹配探索与应用[J].华中科技大学学报,2003(2):23 - 25.