一种基于知网的中文词义消歧算法

张明宝,马 静
(南京航空航天大学 经济与管理学院 信息管理与电子商务系,江苏 南京 210016)

摘 要:词义消歧对自然语言处理领域许多问题的研究具有重要的理论和实践价值。针对该问题,提出了一种基于知网的中文词义消歧算法。为了考虑上下文词汇对词义消歧的不同影响,以语义相似度计算为基础,设计了三种语义联系强度计算方法,并且制定了四条词义消歧规则,依此实现中文词义消歧。实验数据显示该方法可获得65%左右的召回率和75%左右的准确率。

关键词:词义消歧; 语义相似度; 知网
中图分类号:TP391 **文献标识码**:A **文章编号**:1673-629X(2009)02-0009-03

An Approach to Chinese Word Sense Disambiguation Based on HowNet

ZHANG Ming-bao, MA Jing
(Department of Information Management and Electron Business, School of Economy and Management, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China)

Abstract: The automatic disambiguation of word senses has great theoretical and practical significance in many fields of natural language processing. Presents an approach to Chinese word sense disambiguation based on HowNet. In order to take into account different effects of context to word sense disambiguation, three methods of calculating sense relation strength and four related rules are designed based on semantic similarity computing. The recall/ accuracy rate of experiment are respective about 65% and 75%.

Key words: word sense disambiguation; word sense similarity; HowNet

0 引 言

相关研究已经表明词义消歧对机器翻译、信息检索、文本分析、自动文摘、知识挖掘等多方面[1,2]都具有十分重要的作用。现阶段词义消歧研究的方法分为基于词典的方法和基于语料库的方法两大类。基于词典的词义消歧研究强调从各类字典、词典中获得词汇语义间的关系,通过将这些关系量化来辅助词义消歧[3,4]。基于语料库的词义消歧研究强调从语料库中学习自然语言的语言规则,以此来实现语义消歧[5,6]。

考虑到目前中文语料库的研究才刚刚起步,而中文知识词典在中文词义消歧领域已经积累了一定的成果,提出了一种基于知网的中文词义消歧算法,描述了该算法过程及其实验结果,最后对该算法作了性能分析。

收稿日期:2008-05-09
基金项目:国防技术基础项目(1009-234039)
作者简介:张明宝(1973-),男,江苏南京人,博士,副教授,主要从事企业集成、信息检索、知识挖掘等方面的研究。

1 知网简介

知网是一个以汉语和英语的词语所代表的概念为描述对象,以揭示概念与概念之间以及概念所具有的属性之间的关系为基本内容的常识知识库。它不仅对每个词标注了语义、词性,而且标注了词所属的概念、概念与其内部属性之间的联系以及概念之间的联系。知网已经成为中文自然语言处理研究领域最重要的资源之一。

2 基于知网的中文词义消歧算法

文中以知网为基础提出了一种基于语义相似度计算的中文词义消歧算法。该算法基本思想:计算待消歧词汇的各词义与该词汇上下文各词汇词义的语义相似度,根据语义相似度值所反映出来的词义之间的关联关系来实现多义词在特定上下文中的词义鉴别。

2.1 基于知网的语义相似度计算

目前已经提出多种中文语义相似度计算方法,其中刘群等提出的利用知网计算词汇语义相似度的方法具有较大实用价值[7]。文中工作采用该方法计算中文

词汇语义相似度。

2.2 三类语义联系强度

根据词汇语义与其上下文词汇语义之间普遍存在的几类关系，首先定义三类语义联系强度，利用这三类语义联系强度来量化待消歧词汇与上下文词汇词义间的关系，作为词义消歧的计算依据。设待消歧词汇 W 有 n 个语义 $k_1,k_2,\cdots,k_i,\cdots,k_n(n\geqslant 2,1\leqslant i\leqslant n)$。

定义 1：取待消歧词汇所在句子中前一个实词 W_1 和后一个实词 W_2。若待消歧词汇处于句首，则只取 W_2，若待消歧词汇处于句末，则只取 W_1。设这所取实词共有 m 个词义 $r_1,r_2,\cdots,r_j,\cdots,r_m$，$\underset{1\leqslant i\leqslant n,1\leqslant j\leqslant m}{\mathrm{Sim}}(k_i,r_j)$ 为 k_i 和 r_j 的语义相似度值，则待消歧词汇 W 的词义 k_i 在该上下文中的第一类语义联系强度 S_{1i} 为 $\underset{1\leqslant j\leqslant m}{\mathrm{Max}}(\mathrm{Sim}(k_i,r_j))$。

第一类语义联系强度实际上是指待消歧词汇每一词义与其最接近的 2 个上下文词汇的所有词义的语义相似度最大值，它反映了待消歧词汇前一个实词与后一个实词对其词义鉴别的指征作用。

定义 2：取待消歧词汇所在段落中的所有 m 个实词为 $W_1,W_2,\cdots,W_j,\cdots,W_m(1\leqslant j\leqslant m)$。这些实词的词义为 $r_1^1,r_1^2,\cdots,r_1^p,r_2^1,r_2^2,\cdots,r_2^q,\cdots,r_j^g,\cdots,r_m^1,r_m^2,\cdots,r_m^t$。设 $\mathrm{Sim}(k_i,r_j^g)$ 为 k_i 和 r_j^g 的语义相似度值，$\underset{1\leqslant g\leqslant y}{\mathrm{Max}}(\mathrm{Sim}(k_i,r_j^g))$ 为 k_i 和 W_j 的所有词义的语义相似度的最大值，其中 y 为 W_j 的词义个数。则待消歧词汇 W 的词义 k_i 在该上下文中的第二类语义联系强度 $S_{2i}=\sum_{1\leqslant j\leqslant m}\theta_{ij}$，其中

$$\theta_{ij}=\begin{cases}1 & \text{if}\ \exists j_0\exists g_0((j=j_0)\wedge(\mathrm{Sim}(k_i,r_{j_0}^{g_0})=\\ & \quad \underset{1\leqslant j\leqslant m,1\leqslant g\leqslant y}{\mathrm{Max}}(\mathrm{Sim}(k_i,r_j^g))))\\ 0 & \text{else}\end{cases}$$

文章的段落往往具有一个明确的主题，这一主题是由段落中的所有实词所表现的。第二类语义联系强度计算了待消歧词汇的每一词义与段落中所有上下文词汇中具有最强语义相似度的词义数目，它反映了待消歧词汇的每一词义与上下文主题之间的联系倾向性。

定义 3：取待消歧词汇所在句子的前后 χ 句中的 m 个实词为 $W_1,W_2,\cdots,W_j,\cdots,W_m(1\leqslant j\leqslant m)$。这些实词的词义为 $r_1^1,r_1^2,\cdots,r_1^p,r_2^1,r_2^2,\cdots,r_2^q,\cdots,r_j^g,\cdots,r_m^1,r_m^2,\cdots,r_m^t$。设 $\mathrm{Sim}(k_i,r_j^g)$ 为 k_i 和 r_j^g 的语义相似度值，$\underset{1\leqslant g\leqslant y}{\mathrm{Max}}(\mathrm{Sim}(k_i,r_j^g))$ 为 k_i 和 W_j 的所有词义的语义相似度的最大值，其中 y 为 W_j 的词义个数。则待消歧词汇 W 的词义 k_i 在该上下文中的第三类语义联系强度 $S_{3i}=\sum_{1\leqslant j\leqslant m}\eta_{ij}$，其中

$$\eta_{ij}=\begin{cases}1 & \text{if}\ \exists j_0\exists g_0((j=j_0)\wedge(\mathrm{Sim}(k_i,r_{j_0}^{g_0})=\\ & \quad \underset{1\leqslant j\leqslant m,1\leqslant g\leqslant y}{\mathrm{Max}}(\mathrm{Sim}(k_i,r_j^g)))\wedge(\mathrm{Sim}(k_i,r_{j_0}^{g_0})\geqslant\gamma))\\ 0 & \text{else}\end{cases}$$

其中 γ 为阈值。

第三类语义联系强度计算在一定上下文范围内待消歧词汇的各词义与上下文各词汇词义的语义相似度最大值大于某一固定阈值的数目。通过调整 χ 和 γ 这两个参数值可以在最短的上下文范围内寻找最强的语义联系。关于 χ 和 γ 这两个参数值的确定可通过程序试凑获得。

2.3 基于三类语义联系强度的消歧规则

在词义消歧过程中，这三类语义联系强度本身对词义鉴别的指示性会随文档结构、内容等的变化而变化，但是这三类语义联系强度的综合应用会获得较强的对词义鉴别的指示性。为了综合应用这三类语义联系强度，定义了如下词义消歧规则。

已知待消歧词汇的 n 个词义为 $k_1,k_2,\cdots,k_i,\cdots,k_n(n\geqslant 2,1\leqslant i\leqslant n)$，其对应的第一类语义联系强度值为 $S_{11},S_{12},\cdots,S_{1i},\cdots,S_{1n}$，其对应的第二类语义联系强度值为 $S_{21},S_{22},\cdots,S_{2i},\cdots,S_{2n}$，其对应的第三类语义联系强度值为 $S_{31},S_{32},\cdots,S_{3i},\cdots,S_{3n}$。

规则 1：设 $S_{1i}=\underset{1\leqslant t\leqslant n}{\mathrm{Max}}S_{1t}$，$\Delta_1=\underset{1\leqslant j\leqslant n,i\neq j}{\mathrm{Min}}(S_{1i}-S_{1j})$，如果 $\frac{\Delta_1}{S_{1i}}\geqslant\gamma_{10}$，则按照第一类语义联系强度计算该待消歧词汇在该上下文中可能的词义为 k_i。

规则 2：设 $S_{2i}=\underset{1\leqslant t\leqslant n}{\mathrm{Max}}S_{2t}$，$\Delta_2=\underset{1\leqslant j\leqslant n,i\neq j}{\mathrm{Min}}(S_{2i}-S_{2j})$，如果 $\frac{S_{2i}}{\sum_{1\leqslant t\leqslant n}S_{2t}}\geqslant\gamma_{20}$ 且 $\frac{\Delta_2}{S_{2i}}\geqslant\gamma_{21}$，则按照第二类语义联系强度计算该多义词在该上下文中可能的词义为 k_i。

规则 3：设 $S_{3i}=\underset{1\leqslant t\leqslant n}{\mathrm{Max}}S_{3t}$，$\Delta_3=\underset{1\leqslant j\leqslant n,i\neq j}{\mathrm{Min}}(S_{3i}-S_{3j})$，如果 $S_{3i}\geqslant\gamma_{30}$ 且 $\Delta_3\geqslant\gamma_{31}$，则按照第三类语义联系强度计算该待消歧词汇在该上下文中可能的词义为 k_i。

规则 4：如果 k_i 满足规则 1～3 中任意两个或两个以上的规则，则按照三类语义联系强度计算的综合结果，k_i 为该待消歧词汇在该上下文中的确定词义。

上述四条词义消歧规则的有效性完全依赖于 γ_{10}，γ_{20}，γ_{21}，γ_{30}，γ_{31} 这几个参数的合理设置，可通过实验确定。

2.4 基于知网的中文词义消歧算法过程

根据前述四条词义消歧规则，设计了如下的中文词义消歧算法过程：

步骤一，对目标文档进行分词、词性标注处理，取

所有实词进行运算;

步骤二,根据待消歧词汇的词性查询知网,如果该词性的词义只有一个,则按词性标注即可确定其词义;如果该词性的词义多于 1 个,则取出具有相同词性标注的所有词义,作为消歧运算的输入。

步骤三,分别计算待消歧词汇各词义的第一类语义联系强度值、第二类语义联系强度值和第三类语义联系强度值,按照规则 1~3 分别计算待消歧词汇可能的语义。

步骤四,按照规则 4 进行词义消歧,如果识别词义成功,则算法结束;否则通知用户无法识别待消歧词汇词义,算法结束。

3 实验及结果分析

3.1 实验方法

根据文中的词义消歧算法开发了 WSDTool v1.0 系统,该系统使用中科院汉语词法分析系统 ICTCLAS3.0 作为分词和词性标注工具,使用刘群等人开发的软件包作为词汇语义相似度计算工具,使用知网 2.0 作为语义词典。

由于现有的汉语词义消歧公开测评,如 SEMEVAL-2007 task 5 等主要用于基于语料库的词义消歧算法的测评,其测试语料以句子为单位而导致上下文太短,这些都无法满足本算法的需要。所以从人民日报标注语料库和国家语委现代汉语语料库抽取文档来进行实验。

参照信息检索的评价方法,使用召回率和准确率进行实验结果的评价,其计算方法如下:

$$召回率=\frac{进行正确语义消歧的词汇数目}{测试集中所有待消歧词汇数目}\times 100\%$$

$$准确率=\frac{进行正确语义消歧的词汇数目}{测试集中所有待消歧词汇数目-无法进行语义消歧的待消歧词汇数目}\times 100\%$$

3.2 实验数据

前述所设计的四条词义消歧规则中,γ_{10},γ_{20},γ_{21},γ_{30},γ_{31} 这几个参数的合理性对词义消歧的影响非常大,通过实验选择各参数如下:

$$\gamma_{10}=0.8,\gamma_{20}=0.7,\gamma_{21}=0.7,\gamma_{30}=4,\gamma_{31}=2$$

表 1 所示为对选定多义词进行词义消歧的部分结果。表 2 所示为对从人民日报语料库中选定的完整文档进行词义消歧的部分结果。实验数据表明,语义消歧算法可获得平均 65% 左右的召回率和平均 75% 左右的准确率。从召回率和准确率这两个指标来综合衡量,使用本算法可以获得较好的词义消歧效果。

表 1　对选定多义词的词义消歧结果

等消歧词汇	数目	语义 1			语义 2			平均召回率	平均准确率
		文档数	召回率	准确率	文档数	召回率	准确率		
师长	815	425	77%	85%	390	84%	88%	79%	86%
骨干	715	479	59%	63%	236	73%	83%	60%	65%
病毒	567	320	76%	84%	147	58%	57%	63%	78%
对象	635	198	51%	69%	437	57%	75%	54%	71%
油茶	910	276	53%	52%	634	77%	90%	67%	78%
杜鹃	640	350	71%	88%	290	61%	75%	66%	82%
竹叶青	856	400	67%	84%	456	83%	93%	75%	89%
指针	845	276	55%	62%	569	84%	92%	70%	79%
单位	768	327	42%	51%	441	63%	79%	52%	68%
舌头	667	287	70%	81%	380	66%	73%	67%	76%

表 2　对选定文档的词义消歧结果

文档标号	多义词数目	消歧召回率	消歧准确率
19980101-01-001	164	70%	77%
19980101-01-002	152	65%	83%
19980101-01-003	38	70%	79%
19980101-01-004	88	60%	80%
19980101-02-001	66	58%	74%
19980101-02-003	39	50%	58%
19980101-02-006	41	64%	83%
19980101-02-007	103	62%	82%
19980101-02-008	34	59%	77%
19980101-03-001	65	76%	83%

3.3 讨　论

从实验数据来看,影响本算法准确性的因素主要有如下几个方面:

首先,知网收录词汇的不完备以及词汇义项描述的不合理是影响词义鉴别的一个重要原因。对于多义词"病毒",算法对其语义 1-"N bacteria|微生物"的鉴别能力较强,而对其语义 2-"N software|软件,* damage|损害,# software|软件"的鉴别能力很弱。这是由于语义 2 的上下文词汇多是一些与计算机相关的专业术语,知网没有收录,使得这些词汇无法为语义 2 提供指示作用。

其次,所使用的语义相似度值计算方法对知网的义项解释结构依赖性很大。知网的词汇义项描述中,第一义原往往指明了该义项的类别,而我们所使用的语义相似度计算方法中第一义原的影响最为显著。实验发现:如果多义词的义项解释中具有相同第一义原,那么对这些多义词的语义消歧的召回率和准确率都会有一定程度的降低。譬如:"扩张"的语义 1-"V enlarge|扩大"与语义 2-"V enlarge|扩大,medical|医"就具有相同的第一义原,对该词的语义消歧效果就明显的变差。

第三,某些文档中待消歧词汇的上下文对词义鉴别缺乏明显的指示作用,即文档本身不具备词义消歧的基础。对这类上下文使用本算法的词义消歧效果变

(下转第 15 页)

framenumber 域是 11 位的，所以最大为 0x7ff，如果大于的话计数应该归 0，否则加 1。

2.2.6　CRC5 校验码生成模块

由于 MPC8272 的 USB 主控制器不能自动产生 CRC5 校验码，所以也要以软件的方式来为数据提供 CRC5 校验码。CRC5 校验主要应用于 SOF 帧的帧数 framenumber 域校验和对令牌包中的 ADDR 和 ENDPOINT 域校验。根据 USB 协议的规定，SOF 帧和令牌包的格式如图 4 和图 5 所示：其中 framenumber 和 ADDR，ENDPOINT 的位数都为 11 位，所以对于 CRC5 校验码的设计就是利用 CRC5 校验算法对所有可能的 2^11 = 1024 种的数据都计算一次 CRC5 校验码，并保存在一个大小为 1024 的数组 crc5_table[ ]中。要使用时就通过查这个数组来得出相应的校验码来直接填入对应的 CRC5 校验码域。

域	PID	Frame Number	CRC5
位数	8	11	5

图 4　SOF 帧格式图

域	PID	ADDR	ENDP	CRC5
位数	8	7	4	5

图 5　令牌包格式图

2.3　调试和测试主机控制器驱动

实验采用一台主机和 MPC8272 目标板通过串口进行连接，先将经过交叉编译的内核镜像文件载入目标板中，启动内核后将 USB 主机控制器驱动的可执行文件也载入目标板上，再通过 Linux 操作系统 insmod 命令加载它；然后插入 USB 设备，驱动程序可以识别设备并可以正常的读写，验证了该 USB 主机控制器的驱动是可行的。

3　结束语

介绍了 USB 协议以及在 Linux 下 USB 系统的结构，并在 PowerPC 架构下对 USB 主机控制器驱动进行设计和实现，对同类的嵌入式 USB 主机控制器的设计提供了一定的借鉴作用。由于 USB 接口简单易用以及普遍性，本设计对采用 MPC8272 芯片的产品提供了很大的扩展功能作用。

参考文献：

[1] 刘　森. 嵌入式系统接口设计与 Linux 驱动程序开发[M]. 北京：北京航空航天大学出版社，2006.

[2] Rubini A. Linux Device Drivers[M]. 3rd Edition. [s. l.]: O'Reilly, 2005.

[3] 胡晓军，张爱成. USB 接口开发技术[M]. 西安：西安电子科技大学出版社，2005.

[4] 周立功. USB2.0 与 OTG 规范及开发指南[M]. 北京：北京航空航天大学出版社，2004.

[5] 刘胜军，高　济. 嵌入式 Linux 下 USB 主控制器驱动的研究[J]. 计算机应用，2006(3)：25－27.

[6] 唐祎玲，白似雪. 基于嵌入式 Linux 的 USB 主控制器驱动设计[J]. 微计算机信息，2007，23(11－2)：8－10.

(上接第 11 页)

差。譬如，“单位”的词义 1－“N attribute|属性，amount|多少，&entity|实体”的词义消歧的召回率和准确率都较低。分析实验语料发现，该词义与其上下文词汇词义的相似度值普遍偏小，这就使得该词义很难被鉴别出来。

4　结束语

文中提出以知网作为知识源，通过计算待消歧词汇与上下文词汇的语义相似度来分别计算第一、第二和第三语义联系强度，然后依据这三类语义联系强度的语义指向性按照制定的词义消歧的规则来进行词义的鉴别。该算法综合考虑了文档结构、上下文窗口、知网结构以及语义相似度计算方法等方面对词义消歧的影响。实验结果表明该算法是可行的。目前存在的问题主要来自于知网词汇完备性、知网义项描述结构和文档上下文的弱指向性这三个方面，这是下一步工作的主要内容。

参考文献：

[1] Ide N, Veronis J. Introduction to the special issue on word sense disambiguation: the state of the art[J]. Computational Linguistics, 1998, 24(1): 1－40.

[2] Schutze H, Pedersen J. Information retrieval based on word senses[EB/OL]. 2007－12. http://sholar.google.com.cn/.

[3] Li X, Szpakowicz S, Matwin S. A WordNet－based algorithm for word sense disambiguation[EB/OL]. 2007－12. http://sholar.google.com.cn/.

[4] Agirre E, Rigau G. A proposal for word sense disambiguation using conceptual distance[EB/OL]. 1995. http://citeseer.ist.psu.edu.

[5] Leacock C, Chodorow M, Miller G A. Using Corpus Statistics and WordNet Relations for Sense Identification[J]. Computational Linguistics, 1998, 24(1): 147－166.

[6] Leacock C,. Corpus－based statistical sense resolution[EB/OL]. 2007－12. http://sholar.google.com.cn/.

[7] 刘　群，李素建. 基于《知网》的词汇语义相似度计算[EB/OL]. 2007－09. http://www.google.com.