面向搜索引擎的自然语言理解的设计与实现*

钱 兵, 王永成, 高 凯
(上海交通大学 计算机系, 上海 200030)

摘 要: 提出了一种针对智能检索的自然语言理解的实现模型。该模型通过句模分析、分词和概念扩展的方法来理解问句,在一定程度上提高理解自然语言的能力。详细介绍了其系统架构、实现思想和原理。最后通过一系列的实例来对普通搜索引擎和加载了本模型的搜索引擎进行测试。实验结果表明,提出的模型能有效地分析自然语言提问,提高信息检索的准确性和智能性。

关键词: 自然语言理解; 句模分析; 智能检索; 分词
中图法分类号: TP311 文献标识码: A 文章编号: 1001-3695(2006)12-0260-03

Design and Implementation of Natural Language Understanding for Search Engine

QIAN Bing, WANG Yong-cheng, GAO Kai
(*Dept. of Computer Science & Engineering, Shanghai Jiaotong University, Shanghai* 200030, *China*)

Abstract: Natural Language Understanding (NLU) has been one of the highlights in computer science, and the paper proposes a model which can implement NLU for the intelligent search engine, it understands the natural language by sentence model analysis, word segment and concept, so it can increase the ability of understanding the natural language. Then we introduce the system architecture, idea, and theory in detail. Finally, we compare the normal search engine and another which based on our model by some concrete sentences, the results show that the model can increase the precision and the intelligence of information retrieval.

Key words: Natural Language Understanding; Sentence Model Analysis; Intelligent Search; Word Segment

1 引言

随着计算机与网络的迅速发展,当今社会已经走向信息化,人们越来越不满足于以机械的指令来与计算机交互,而是更强烈地希望使用自己熟悉的自然语言来与计算机进行交流。因此,自然语言理解和处理已经成为当前计算机科学中的一个重点研究方向。它所要研究的是如何能让计算机接受和理解人们日常所使用的语言,使得计算机可以做到人类自然语言与自身机械指令之间的相互转换,建立起一种人与机器密切而友好的关系,使之能进行高度的信息传递和认知活动。

作为自然语言应用的领域之一,搜索引擎的人机界面的友好程度是衡量搜索引擎性能的重要指标之一。自然语言由于更加贴近人类的生活习惯,使检索系统与用户的交流更为人性化。多数用户均不习惯使用复杂的布尔逻辑表达式来表述自己的需求,而更倾向于用自然语言进行检索。目前的普通搜索引擎基本上均只是使用关键词技术,每次搜索时按照关键词进行匹配,因此返回的大量信息中有很大一部分不是用户需要的,往往导致用户陷入信息泛滥的沼泽。基于自然语言理解技术的智能搜索引擎可以依靠自然语言理解技术对知识的理解和处理能力,将信息检索从目前基于关键词的层面提高到基于知识(或概念)的层面,从而更大程度地了解用户的信息需求,获得更易用、更准确、更智能的搜索结果。摆脱了传统搜索引擎基于关键词的束缚,并能提供相关的、有参考价值的其他内容,因而具有信息服务的智能化、人性化特征。因此,将自然语言处理技术引入检索系统的研究也是搜索引擎技术的一个重要方向,同时也是智能化搜索引擎的一个难点[1]。

2 自然语言理解与处理的现状

现阶段的自然语言理解技术有两大研究方向,即基于规则的分析方法(即所谓的理性主义)和针对大规模语料库的分析法(即所谓的经验主义)。这两种方法都有自己的优点,但也存在一些薄弱的地方,主要体现在:

(1) 前者的核心是分析句子结构,利用规则自顶向下或自底向上的句法树生成过程来分析。随着这些年的发展,具有代表性的有:词汇功能语法(LFG)、功能合一语法(FUG)和广义短语结构语法(GPSG)等。这些方法基本上掌握了单个句子的分析技术,但在处理大规模的非受限真实文本时很难覆盖全面,特别是对于整个段落或篇章的理解还无从下手。

(2) 后者充分利用计算机的高速处理能力和海量存储,收集大量相关的文本建立语料库,然后在语料中标注各种记号,标注的内容包括每个词的词性、语义项、短语结构、句型和句间关系等。随着标注程度的加深语料库逐渐熟化,成为一个分布的统计意义上的知识源。利用这个知识源可以进行许多语言

收稿日期: 2005-10-25; 修返日期: 2005-12-28
基金项目: 上海市信息化专项基金资助项目(050211)

分析工作,如根据从已标注语料中总结出的频度规律可以给新文本逐词标注词性,划分句子成分等。但语料库提供的知识是用概率统计表示的,因而常常会出现答非所问的情况,并且语料库需要耗费大量的时间和资源去建立、维护,成本很高。

这两种方法均有自己的特点,但也存在着以上介绍的弊端,而两种方法的融合才是当前自然语言处理发展的趋势。

3　系统主要架构

本文的研究是应用于搜索引擎上的。用户查询所用的自然语言一般以单个提问句为主,相对来说句式较少,意思明确。因此以使用基于规则的分析方法为基础,中间词法分析采用基于语料库的分析法,形成互补,逐步积累逐步探讨,进而扩展到其他句式,从而达到从点到面的目的。系统设计的主要思想为:首先,对用户提交的以自然语言表述的问题进行提问句式分析,去掉其中的提问特征词;然后对剩余的语句片断进行分词处理;分词完以后再去掉其中无意义的辅助词;最后进行概念扩展等处理,利用提取出的核心词及其扩展后的概念进行查询。

本系统用户采用自然语言的方式与系统进行交互,系统接收到用自然语言描述的问题后,根据图 1 所示的流程对问题进行处理,最后将查询到的结果返回给用户。

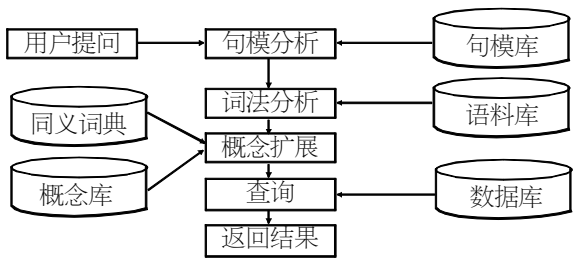


图 1　系统总体架构

总体上分为以下步骤:

(1) 句模分析。接收用户以自然语言表述的问题,在句模库的支持下判断该问题属于哪种提问句式,然后根据提问特征词对其进行切分,去掉提问特征词,得到语句片断(详见 4.1 节)。

(2) 词法分析。根据语料库对所得到的语句片段进行分词处理,去除一些无意义的仅用于构成句子辅助词,抽取出核心语义项,形成查询主题突出、概念明确的短语集合(详见4.2 节)。

(3) 概念扩展。从同义词典和概念库中找出核心语义项对应的各种同义词和概念进行扩展,并且可以根据提问句的类型加入一些对应辅助查找的词(详见 4.3 节)。

4　系统实现主要步骤

4.1　句模分析

当用户提交了以自然语言表述的问题后,根据句模库里收集的各种句模,将问题与句模进行模式匹配,判断该问题属于哪类问题,然后根据句模的定义,舍弃掉其中的疑问词和问句特征词,将句子切分开,得到语句片断。

4.1.1　句模定义

由于本系统是针对提问句式来进行设计的,而中文文法中提问类型是有限的,因此笔者收集了最常用的提问句型,分析并总结其句模和规律,然后使用巴柯斯范式(Backus Normal Form)来定义它们的文法,放在句模库中。

例:小明什么时候上大学?

可以定义其文法为

<询问时间语义块>::=(<事件>)<提问时间特征词>(<事件>)

<提问时间特征词>::=何时|什么时候|哪天|几月几日|…

句模库的大致数据结构如表 1 所示。

表 1　句模库数据结构

编号	名称	提问特征词	常用语法结构
1	提问时间类	什么时候/何时/…	什么时候 + v/何时 + v/…
2	提问定义类	是什么/何为/什么是/…	n + 是什么/何为 + n/什么是 + n/…
3	提问数量类	多少/哪些/	多少 + n/哪些 + n
…	…	…	…

4.1.2　分析问句

当接收到用户的问题后,首先根据提问的特征词使用正则表达式对句模库中的句模进行匹配,分析出该问题属于何种句模。

对问句进行分析的意义在于:首先它对理解用户的提问意图有很大帮助,可以有针对性地进行查询,如在概念扩展阶段可以加入一些针对特定种类查询的特定辅助词(如查询处所的时候可以加上“地处”“位于”等辅助查询关键词);其次通过句模分析后可以去掉句子中的提问特征词(如“什么”“哪里”等),将大大减小下一阶段词法分析中出现歧义的概率。总之,句模分析阶段处理得到的成果,将为后续处理提供关键的帮助。

4.1.3　切分句子

问句分析完成以后,根据句模的定义,舍弃问句中提问特征词,将句子切分开。

例如:“韩正什么时候去法国访问”

首先根据句模的定义进行匹配,根据“什么时候”关键词分析出问题属于“提问时间”句模;然后将问题句中“什么时候”关键词舍弃掉,把句子切分开,这样句模分析后得到的结果是“韩正”“去法国访问”。

4.2　词法分析

每个提问句中均有其核心语义项,核心语义项代表了用户所提问题的实质,也是自然语言理解中的本质任务,而这些核心语义项就隐藏在上一阶段得到的语句片断中。

根据语料库对所得到的语句片段分词,去除一些构成句子所必需的无意义的辅助词,抽取出核心语义项,得到查询主题突出、概念明确的短语集合。

4.2.1　分词

由于句模分析后得到的语句片断是切分后得到的结果,所以它的灵活性和多样性比真实的自然语言文本要差一些,这样当进行词法分析的时候出现歧义的情况较少。传统机械式分词方法主要有正向最大匹配算法(FMM)和逆向最大匹配算法(RMM)[3]。本系统采用两种相结合的双向扫描的方法,这样可以最大限度地来识别歧义字段。扫描后的结果一般会是如下两种:

(1) 当双向扫描后两种分词的结果一致时,认为这次切分为正确的切分。

例如: “韩正”“去法国访问”正向最大匹配后分词的结果为“韩正”“去/法国/访问”,而逆向最大匹配后分词的结果也为“韩正”“去/法国/访问”。这时候就认为这次切分是无歧义、正确的。

(2) 在经历了第一阶段去除提问特征词后,绝大部分切分后都是正常的。但由于汉语的灵活性,还是会出现两次切分得到结果不一样的情况,这时就认为出现了歧义切分。

例如: “研究生物科学的学生”正向最大匹配后分词的结果为“研究生/物/科学/的/学生”,而逆向最大匹配后分词的结果却为“研究/生物/科学/的/学生”,这时候认为这次切分为歧义切分。

出现这种情况时,可以利用词句切分概率对歧义字段进行处理,具体做法如下:

定义 $f(W)$ 代表一个词 W 的词频,它是 W 在语料库中出现的次数/语料库中各词出现的次数之和;

$p(S)$ 代表语句 S 的某一切分形式出现的概率;

则 $p(S) = f(W_1) \times f(W_2) \times f(W_3) \times \cdots \times f(W_t)$,其中 t 为切分完后词的个数。定义 H 为信息熵的值,那么,计算语句切分效率的基本公式为 $H = -\log_2(p(S))$,通过计算得到 H 值,其中 H 值较小的切分即为正确切分。

通过计算和比较之后,认为“研究/生物/科学/的/学生”是正确切分。

4.2.2 去除辅助词

词法分析结束以后得到一系列短语的集合,其中既包含了用户查询目的的核心词汇,但也掺杂了一些构成句子所必需的辅助词,如“的”“地”或一些动词等一系列无意义的词,它们不能明确表示用户要查询的内容,反而会影响查询结果的准确性。

因此首先建立一个过滤词表,其中包括各种组成句子所必需但又无具体意义的字、词;然后根据过滤词表对切分完后的结果进行分析,将提问中出现在过滤词表中的词删去,抽取出核心语义项,形成查询主题突出、概念明确的短语集合。

例如: “韩正”“去/法国/访问”抽取出的核心语义项即为“韩正”“法国”“访问”。

4.3 概念扩展

首先,从同义词典和概念库中找出核心语义项对应的各种同义词和概念进行扩展。

例如: “我国有多少辆自行车”其中的“自行车”可以扩展为“自行车”/“bicycle”/“脚踏车”/“单车”等。

其次,在句模分析阶段中已经知道问题属于哪种类型,这样可以根据提问句的类型加入一些辅助查找词。首先为每种句模建立一个对应的辅助词表,词表中的内容为根据不同的句模确定的不同辅助词。

例如: “中国科技大学在哪里?”可以扩展为“中国科技大学”+“位于”/“地处”等。

5 系统测试结果

为验证系统自然语言理解技术的准确性,笔者作了一个测试,对广泛征集到的 1 700 余条自然语言提问句放入到智能搜索引擎系统中进行测试,然后将搜索得到的结果进行人工验证,测试时采用一个指标: 正确率(Precision)。其计算公式如下:

$$\text{Precision} = a/b \times 100\%$$

其中,a 为搜索结果匹配的提问句个数; b 为所有的提问句个数。

测得的数据如表 2 所示。

表 2 系统测试结果

引擎	普通搜索引擎	智能搜索引擎(本系统)
正确率(Precision)	65%	93%

经过分析以上实验数据,发现该系统对于绝大部分语法比较规范的提问句查询均能得到准确和全面的结果,其中出现不准确的结果一般是由于提问句语法不规范或者是由于句模没有总结完善的原因。但由于系统具有良好的扩展性,可以随时扩展和调整句模库以及对应的各类辅助模块。经过逐步的测试,针对绝大部分日常的提问,该系统都可以做到集精确、全面和智能为一体。

6 结束语

自然语言理解与处理已经日益成为计算机科学领域中的一个热点,具有广阔的研究前景和实用价值。本文结合了传统的基于规则的分析方法和基于大规模语料库的分析法,将自然语言理解和处理技术应用于搜索引擎中,从而提高了检索的精度和智能性。

通过以上研究,发现有以下几个问题需要解决: 首先,如何建立一个更全面的句模库以及完善对应的辅助模块; 其次,如何将在解决提问句方面的经验和启发拓展到更大的领域,而不是仅仅局限于查询的提问句。这些问题均有待于以后进一步的研究。

参考文献:

[1] 王永成. 中文信息处理技术及其基础[M]. 上海: 上海交通大学出版社, 1991. 125-213.

[2] James Allen. Natural Language Understanding(2nd edition)[M]. The Benajmins/Cummings Publishing Company Inc., 2004. 63-92.

[3] 刘群. 汉语词法分析和句法分析技术综述[C]. 第一届学生计算语言学研讨会专题讲座, 2002. 32-37.

[4] 刘群, 李素建. 基于《知网》的词汇语义相似度计算[C]. 第三届汉语词汇语义学研讨会, 2002. 76-80.

[5] 马晓娜, 杨承磊. 基于面向对象的受限自然语言查询系统的设计与实现[J]. 计算机工程与应用, 2005, (10): 165-168.

[6] 欧振猛, 余顺争. 中文分词算法在搜索引擎应用中的研究[J]. 计算机工程与应用, 2000, 36(8): 80-84.

[7] 姚天顺, 朱靖波, 张利, 等. 自然语言理解: 一种让机器懂得人类语言的研究(第 2 版)[M]. 北京: 清华大学出版社, 2002. 212-253.

作者简介:

钱兵(1981-),男,江苏如皋人,硕士研究生,主要研究方向为自然语言理解与处理; 王永成(1939-),男,上海人,教授,博导,主要研究方向为信息检索、中文信息处理、办公自动化及人工智能; 高凯(1968-),男,河北石家庄人,博士研究生,主要研究方向为信息检索。