文章编号: 1001-0645(2004)10-0885-05

基于 ALICE 的汉语自然语言接口

夏 天, 樊孝忠, 刘 林, 骆正华
(北京理工大学 信息科学技术学院计算机科学工程系, 北京 100081)

摘 要: 分析人工智能聊天机器人 ALICE 的知识组织结构和推理机制, 研究 ALICE 在处理汉语时需要解决的问题, 提出利用语义语法扩展知识描述语言 AIML 的表达能力. 采用不确定性推理进行模式搜索并对结果打分择优和答案动态提取, 基于 ALICE 设计实现了一个汉语自然语言接口——CNLIS, 系统结构与具体领域无关. 实验结果表明, 该接口移植方便, 准确率和召回率可分别达到 91.45% 和 91.70%.

关键词: ALICE; 人工智能标记语言; 汉语自然语言接口; 语义语法
中图分类号: TP 391.2 **文献标识码**: A

A Chinese Natural Language Interface Based on ALICE

XIA Tian, FAN Xiao-zhong, LIU Lin, LUO Zheng-hua
(Department of Computer Science and Engineering, School of Information Science and Technology, Beijing Institute of Technology, Beijing 100081, China)

Abstract: The knowledge organization of ALICE and its deduction mechanism are analyzed. Some shortcomings can arise when ALICE is used in processing Chinese. Relevant solutions on extended semantic grammar, pattern-matching algorithm and dynamic answer extraction are proposed. A domain independent Chinese natural language interface CNLIS is implemented based on ALICE. Experiments prove that CNLIS can be easily combined with domain knowledge and achieve 91.45% precision and 91.70% recall.

Key words: artifical linguistic internet computer entity (ALICE); artificial intelligence markup language; Chinese natural language interface; semantic grammar

随着社会的日益信息化, 人们希望能用自然语言与计算机交流, 实现从文字接口、图形接口到自然语言接口的革命.

自然语言人机接口的研究在 20 世纪 60 年代的代表作有 R. Green 的 BASEBALL 系统和 J. Weizenbaum 的 ELIZA 程序, 分析策略依赖于关键词匹配技术[1]. 70 年代 Woods 设计的 LUNAR 系统采用扩充转移网络处理英语语法问题, 对英语的句法和语义做了较深入的分析. 1978 年由 C. Hendrix 提出的 LIFER 系统是一种自然语言通用接口, 利用它建立起了一批自然语言专用接口, 如 LADDER 和 HAWKYE. 90 年代末美国 Lehigh 大学 Richard S. Wallace 博士开发的 ALICE (artificial linguistic internet computer entity) 是一个基于经验的人工智能聊天机器人. 作者研究的 CNLIS (Chinese natural language interface system) 是一个以 ALICE 为基础设计实现的汉语自然语言接口, 在结合金融领域自动问答的应用中, 取得了较

收稿日期: 2003-10-27
作者简介: 夏 天(1978-), 男, 博士生, E-mail: joyxiatian@sohu.com; 樊孝忠(1948-), 男, 教授, 博士生导师.

好的效果[2].

1 ALICE简介

1.1 ALICE与AML

ALICE是一个基于经验的人工智能聊天机器人,在2000年和2001年作为"最像人类的计算机"两次获得Loebner奖[3].ALICE以AML作为知识描述语言,采用了基于实例的推理方法,新知识的获取可在人工指导下进行.在AML中,基本的知识单元由分类(category)组成,每个分类又包括用户输入的问题、ALICE输出的答案和可选上下文环境(optional context)三部分[4].一个简单的分类如下:

```
〈? xml version="1.0" encoding="ISO-8859-1"?〉
〈aiml version="1.0"〉
  〈category〉
    〈pattern〉HOW MANY DAYS * WEEK〈/pattern〉
    〈template〉7 days per week.〈/template〉
  〈/category〉
〈/aiml〉
```

其中〈aiml〉元素指明其中存放的是AML知识;〈category〉代表此处为一个知识实例;〈pattern〉代表用户输入的问句;〈template〉则代表系统应给出的答案;星号表示此处可与1到任意多个单词匹配成功.上例中,如用户输入:"How many days a week?",ALICE将返回"7 days per week".

1.2 ALICE的知识组织和推理

ALICE把所有的知识以树的形式组织到一个Graphmaster对象中,Graphmaster则由一系列称为Nodemapper的节点集组成.每个节点代表模式中的一个单词或统配符,根据它在模式中出现的位置前后相连,构成一棵知识树.类似于计算机中的文件系统,不同模式之间前面相同的部分共享同一父路径,以提高内存利用率.以下面4句为例:①where are you;②where are you going;③where is China;④how are you,其内部组织形式如图1所示,4个句子在树中用9个节点即可表示出来.

根据用户输入的问句在知识树中查找模式的过程是ALICE的推理过程.ALICE在模式搜索时采用了带有回溯的递归过程,首先把用户输入的句子拆分成词,再在知识树中按层次逐个查找.如果与问句相匹配的终端节点包含模板信息,则终止搜索,取出模板内容并进行后处理,返回用户结果.

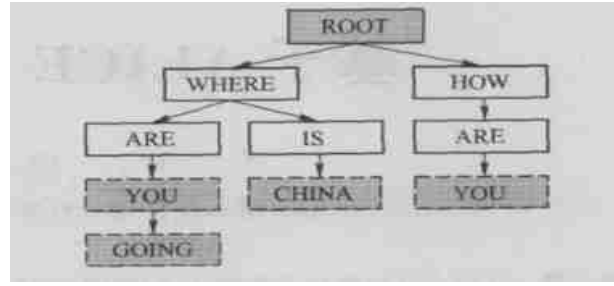

图1 ALICE知识组织例图

Fig. 1 Knowledge organization of ALICE

1.3 ALICE的缺点

ALICE以词作为处理的基本单位,在英语、法语等屈折语对话方面取得了成功,但是在汉语处理方面存在着明显的不足[5].

①ALICE不能处理汉语问答.ALICE以词为其处理的基本单位,英语等西方语言是拼音文字,词之间以空格为分割标志,偶尔出现一些单词缩写和特殊标点符号的处理也较简单,但是以汉字作为最小单位的汉语句子中,词与词之间没有明显的分割标志[6].

②ALICE的推理过程是一个允许有通配符的严格匹配过程,对于语序比较自由,虚词运用较多的汉语来说不合适.

③ALICE的上下文处理能力较弱,标准AML在知识表示时缺乏语义表达能力.

2 基于ALICE的CNLIS

自然语言接口是人工智能领域比较前沿的研究方向之一[7],基于ALICE设计实现的CNLIS,引进了自然语言处理领域中的流行技术,以克服ALICE的不足.

2.1 扩展AML和语义语法

标准AML的语义表达能力较弱,如果一个回答对应多种不同的问法,就要尽量把所有这些问法都枚举出来.以"1加2等于几"为例,在保持基本语序不变的情况下,依然有多种不同说法,如:"1加2等于多少","1加2是几"等等.在编写这类知识时,原有方法是枚举出尽可能多的相似句子,再利用〈srai〉标记跳转到相同的答案模板,编写时很难避免重复或遗漏.为此,扩展了原编写方法,把可能一起出现的成分用花括号括起来,各部分之间的并列成分用竖线隔开;加载知识时,先还原这类模式所有可能的句子组合,再把各句子分别加载.上例可改写

为: 1 加 2{等于 |是}{几 |多少} . 另外, 扩展A ML 还允许一个分类之中出现多个模式, 并对应同一个模板, 该情况在语义上看作对同义句的组织处理 .

语义语法是把大量的语义信息植入到句法描述中去, 用较少的句式来表示众多的输入语句, 以提高语言的概括能力 . 还以"1 加 2 等于几"为例, 标准的A ML 不可能把所有数字相加的模式都加入到知识库中, 而引入语义语法之后, 就可简单地把这类知识表示如下:

〈pattern〉[NUM]加[NUM]{等于 |是}{几 |多少}〈pattern〉
〈template〉
〈calculate〉
〈NUM index = "1"/〉+ 〈NUM index = "2"/〉
〈/calculate〉
〈/template〉

模式中的词(如上例中的"加"或"等于")只能与它们本身和自己的同义词相匹配, 元符(如上例中的[NUM])则可能有多种匹配方式, 这取决于对它们的定义类型 .

① 定义为一个简单的集合 . 例如在模式"[BANK_ CARD]是一种银行卡吗"中, [BANK_ CARD]= {龙卡, 牡丹卡, 葵花卡, …}

② 定义为一个谓词 . 用它去测试输入串是否满足某个条件(例如上例中的[NUM], 就要求该部分字符串必须是一个数字串) .

2.2　上下文处理

AL ICE 本身具有一定的上下文处理能力, 它在对话中记录了自己上一句的输出, 并与知识中的〈that〉部分进行比较, 辨别是否是刚谈到的话题, 以便作出正确的回应, 如:

〈category〉
〈pattern〉NO〈/pattern〉
〈that〉OK IW LL STOP TALKING NOW〈/that〉
〈template〉
But you told me to stop talking
〈/template〉
〈/category〉
〈category〉
〈pattern〉NO〈/pattern〉
〈that〉SEE YOU LATER〈/that〉
〈template〉Well then, not! 〈/template〉
〈/category〉

如果上一句用户输入为"Good bye", 系统输出"See you later", 则当用户再输入"no"时, 系统将输出"Well then, not!".

以上方法虽然在一定程度上提高了系统的逼真性, 但对于话语中传递的许多隐含信息却未做处理, CNL IS 采用移动窗口技术始终保持最近的 8 条对话记录, 并根据这些信息处理回指问题, 还原缺省成份, 以进一步增强计算机的理解能力, 例如以下 3 个问句:

① 什么是储值卡?
② 生肖卡呢?
③ 它有什么特点?

当一个问句在正常分析中失败后, CNL IS 尝试作上下文的缺省还原处理 . 由于相似的句法结构往往具有相似的语义, 如果当前问句分析失败, 就拿它与上下文中没有省略的完整句子进行比较 . 如对句②作省略还原处理时, 根据句法和语义分析, "生肖卡"与上句中的"储值卡"成份相同, 都是名词, 而且在语义上都有银行卡之意, "呢"作为语气词对句子的影响很小, 于是当前问句还原为"什么是生肖卡?". 对于句③, 需要把代词还原, 根据最近上下文"什么是生肖卡", 推出其中心词为"生肖卡", 故把当前代词以"生肖卡"替换, 问句还原为"生肖卡有什么特点?".

3　CNL IS 系统体系结构

CNL IS 主要包括输入预处理、知识组织方式、模式搜索和答案提取 4 个模块 . 系统流程如图 2

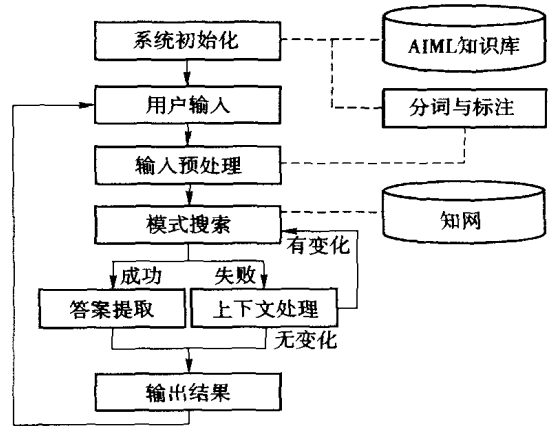

图 2　CNL IS 系统流程图

Fig. 2　Flow sheet of CNL IS

所示.

3.1 输入预处理

预处理的首要功能是对汉语问句进行分词和词性标注.CNLIS采用逆向全切分分词算法,依据词典进行切分,基于词频和词分级加权评估,选取最优的切分结果,并在此基础上进一步对地名、人名、常见的缩略语进行自动识别.对词性标注则采用规则和统计相结合的方法处理.

输入处理还包括对特殊字符的替换处理,例如,把"俺"换成"我",把"it's"换成"it is"等.利用该模块,把用户的问句规范化.

3.2 CNLIS的知识组织方式

为提高速度,充分利用空间,CNLIS借鉴了ALICE的知识组织方式,采用节点复用技术,把所有的知识以树的形式组织到Graphmaster对象中.假设n为知识树中的一个节点,w为一个词,则$G(n, w)$要么尚未定义,要么返回n的后继节点m的值,$S_n=\{w: m \mid G(n, w)=m\}$则定义为以$n$为直接父节点的所有子节点值的集合.例如,设$r$为知识树的根节点,$S_r$就表示为所有模式中第1个词的集合.

在知识加载时,系统先对模式分词,然后按顺序把词从根节点r到末节点t依此存入树中,同时把模板信息记录在末节点t之中.设$w_1, \cdots, w_k$为一模式的连续k个词,要把该模式插入树中,系统首先验证$m=G(r, w_1)$是否已定义,如果已经定义,则以递归的方式继续把$w_2, \cdots, w_k$插入以m为根的子树之中,直到$G(n, w_i)$返回未定义时为止.此时,Graphmaster就创建一个新节点$m=G(n, w_i)$,插入到知识树中,并把余下的词串$w_{i+1}, \cdots, w_k$以同样的方式继续处理,直至结束.

3.3 模式搜索

CNLIS的模式搜索过程是一个带有剪枝的递归过程,大部分情况下匹配结果并不唯一,因此CNLIS利用打分机制从结果中进行择优.设$w_1, \cdots, w_k$为用户输入的问句,在Graphmaster中匹配时,首先调用MatchNet(r, 1)函数,在这里,r表示根节点,1表示问句中单词的位置序号,算法如下:

```
boolean MatchNet(node, index)
{
  result= false;
  if(index> word_count){
    return true;
  }else{
    for each c in (S_n= {w:m |G(node, w)= m}){
      mark= GetMark(c, w index);
      if(mark> ThresholdValue){
        if(MatchNet(G(node, c), index+ 1) = = true){
          PushResultNode(G(node, c), mark, pos);
          result= true;
        }}}
    if(CanSkip(w index) and MatchNet(node, index+ 1)){
      result= true;
    }
    return result;
  }
```

下面对算法做进一步描述:

① 用GetMark函数计算两个单词之间的匹配相似度.设w_1为模式中的词,w_2为问句中的词,如$w_1=w_2$,函数值为1.2;如w_1为*,值为1.0;否则,匹配相似度为二者的语义相似度,相似度的计算基于知网[8,9].当相似度计算结果小于给定的阈值时,系统进一步判断w_1与w_2的拼音是否相同,如相同,函数值为0.9.当w_1在知识模式中的相应节点是一个语义节点时,系统将在规则库的支持下,调用逻辑匹配与推理模块计算匹配结果,其它情况下,函数值为0.

② 缺省阈值(threshold value)为0.8,因此,若GetMark< 0.8,算法将不再搜索以当前节点为根节点的子树.

③ 问句中的当前处理词在某些情况下可以忽略,如助词、语气词和部分形容词,在没有搜索到更合适的模式时,算法将跳过这类词继续处理.

④ 算法把中间结果节点都压入一个栈中,并记录节点的位置、匹配的分数(即GetMark的值)和其他必要信息.算法完成之后,Graphmaster首先从栈中还原出所有满足条件的模式,然后再以每条模式对应的分数总和除以该模式所拥有的单词数目计算均值,从中挑选出均值最大者作为匹配结果.

3.4 答案提取

CNLIS在模板中使用了大量特殊标记,最终结果将根据这些特殊标记和相关参数动态提取.例如

用于数学计算的〈calculate〉, 获取领域信息的〈domain〉等. 系统在取出模板后, 先解析其中的特殊标记, 再把结果返给用户. 利用这些标记, CNL IS 可以方便地对应用领域进行扩充和移植.

4　实验结果分析

北京理工大学NLP 实验室在金融领域自动问答系统的设计与实现中, 采用了CNL IS 与领域问答提取模块相结合的策略. 用户输入的问句首先由CNL IS 处理, 当CNL IS 没有输出内容或者输出内容为一领域问句向量时, 再交给领域问答提取模块处理. 针对2000 个问句的最新封闭测试集, 结果如表1 所示.

表1　CNL IS 实验结果

Tab 1　Experimental results of CNL IS

CNL IS 输出形式	CNL IS 处理条数	满意条数	领域处理条数	满意率/%
问句向量	1 672	1 672	1 672	100.00
直接答案	162	157		96.91
无结果	166		139	83.73

其中, CNL IS 的输出形式可能是最终答案, 也可能是领域问句向量. 如果是问句向量, 还需要把它提交给领域模块解析, 如果CNL IS 未找到匹配模式, 就由领域模块直接提取答案.

实验结果表明, 83.6% 的问句经CNL IS 处理之后, 可以去掉其中无关紧要的成分, 解析出具体的问句向量, 进而由领域模块提取相关答案.

5　结束语

基于AL ICE 的CNL IS, 作为一个结构通用、与专业知识结合方便的汉语自然语言接口, 在网络教学、呼叫中心、FAQ 查询等诸多领域中都有广阔的应用前景, 并在与金融领域问答系统的结合测试中, 取得了令人满意的效果. 但是, 要让CNL IS 更加成熟、稳定, 面向实际应用, 仍然还有大量工作要做, 如知识标记集合的制定与标记语言的规范化和标准化、答案提取中的逻辑推理、知识自动获取、包括常识知识库、领域知识库和规则库在内的知识库建设等, 都需要进一步深入研究.

参考文献:

[1] 刘开瑛, 郭炳炎. 自然语言处理[M]. 北京: 科学出版社, 1991.
Liu Kaiying, Guo Bingyan. Natural language processing [M]. Beijing: Science Press, 1991. (in Chinese)

[2] Li Hongqiao, Fan Xiaozhong, LI Liangfu, et al. The study and implementation of finance-domain Chinese automatic question-answering system: FAQAS [A]. Advances in Computation of Oriental Languages[C]. Beijing: Tsinghua University Press, 2003. 483- 489.

[3] Wallace R S. The anatomy of AL ICE[EB/OL]. http: //www. alicebot. org/anatomy. html, 2001-03-25/2003-07-25.

[4] Wallace R S. AIML overview [EB/OL]. http: //www. pandorabots. com/pandora/pics/wallaceaimltutorial. html, 2001-03-25/2003-07-25.

[5] 夏　天, 樊孝忠, 刘　林. AL ICE 机理分析与应用研究[J]. 计算机应用, 2003, 23(9): 1- 5.
Xia Tian, Fan Xiaozhong, Liu Lin. AL ICE mechanism analysis and application study [J]. Computer Applications, 2003, 23(9): 1- 5. (in Chinese)

[6] 刘　颖. 计算语言学[M]. 北京: 清华大学出版社, 2002.
Liu Ying. Computing linguistics [M]. Beijing: Tsinghua University Press, 2002. (in Chinese)

[7] 姚天顺, 朱靖波, 杨　莹等. 自然语言理解——一种让计算机懂得人类语言的研究[M]. 北京: 清华大学出版社, 2002.
Yao Tianshun, Zhu Jingbo, Yang Ying, et al. Natural language understanding—— A research of making computers understand human language [M]. Beijing: Tsinghua University Press, 2002. (in Chinese)

[8] Li Sujian, Zhang Jian, Huang Xiong, et al. Semantic computation in a Chinese question-answering system [J]. Journal of Computer Science and Technology, 2002, 17(6): 933- 939.

[9] 董振东, 董　强. 知网[EB/OL]. http: //www. keenage. com/download/hownetsystem. exe, 2001-11-02/2003-07-27.
Dong Zhendong, Dong Qiang. Hownet [EB/OL]. http: //www. keenage. com/download/hownetsystem. exe, 2001-11-02/2003-07-27. (in Chinese)