中 文 信 息 学 报
第 16 卷 第 6 期 JOURNAL OF CHINESE INFORMATION PROCESSING Vol. 16 No. 6

“CAU”词及其知识图分析*

刘小冬[1] 张 蕾[2]

(1. 西北工业大学应用数学系 西安 710072 2. 西北大学计算机科学系 西安 710069)

摘要:专家系统是人工智能研究领域的一个重要研究分支。专家系统主要由两部分组成:知识库和推理机。知识库中的知识主要由“IF—THEN”这样的知识组成。知识图是一种新的知识表示方法。在知识图中,含有“IF—THEN”结构的句子是由起因操作符(causal operator)或起因关系(CAU-relation)表示的。本文挑选了一些具有一定代表性的起因意义的汉语“CAU”操作符,并且基于知识图理论分析了这些操作符,并进行了分类,目的是为专家系统中知识库的建立做准备。

关键词:专家系统;知识图;知识库;起因单词

中图类号:TP391

“CAU”Words and the Analysis by Means of Knowledge Graphs

LIU Xiao-dong[1] ZHANG Lei[2]

(1. Department of Applied Mathematics Northwestern Polytrchnical University Xi'an 710072

2. Department of Computer Science Northwest University Xi'an 710069 China)

Abstract: Expert systems form one of the most important research areas in Artificial Intelligence. The main parts in expert systems are knowledge bases and inference engines. In the knowledge bases the main knowledge is knowledge expressed by “IF-THEN” statements. In knowledge graphs, a new form of knowledge representation, the “IF-THEN” statements are tired up with causal operators (CAU-relations). In this paper, we picked out some Chinese operators with “CAU” meaning, and investigated these operators. The goal is to build knowledge bases in expert systems.

Key words: expert system; knowledge graph; knowledge base; CAU word

一、引言

专家系统是使用知识有效地解决狭窄领域内复杂问题的计算机系统[5]。专家系统主要由两部分组成:知识库和推理机。用于解决问题的、狭窄领域内的知识被储存在知识库里面。知识库的知识主要是由“IF—THEN”这样的知识组成。为了解决不同的狭窄领域内的问题,在不同的知识库中有不同的知识表达方法。知识库中的知识决定了专家系统的功能、有效性。推理机使用知识库中的知识来解决问题。

从专家或文本中自动获取知识的方法是表示知识、并且把知识存储到知识库中的一种方法。因此,在专家系统中获取知识是十分重要的。自动获取知识的技术是建立专家系统的瓶颈问题。研究专家系统自动获取知识的技术是专家系统研究者所关注的焦点问题[6]。

知识图方法是一种新的知识表示方法[2~7]。如何从文本中自动获取具有“IF—THEN”这

* **收稿日期**:2002-5-20

基金项目:航空科学基金项目(01J53079)

作者刘小冬,男,1963 年生,副教授,博士,主要研究方向为自然语言理解、运筹学等. 张蕾,女,1964 年生,副教授,博士,主要研究方向为自然语言理解、人工智能、数据库等.

样结构的知识,知识图理论为我们提供了一条途径。在知识图理论中,具有"IF—THEN"关系的句子,是通过"CAU"关系(起因关系)来表示的。因此,区分、寻找一句话中具有"CAU"关系的单词或词组(操作符,*operator*)和该关系所涉及的概念(操作数,*argument*)是建立专家系统的关键。操作数是自封闭的单词或词组,而操作符是一个依赖于一个或多个操作数的单词或词组。操作数被看作是概念,而操作符被认为是概念之间的关系。

知识图的基本思想是从文本中抽取概念之间的各种各样的关系,使用节点表示概念、连接节点之间的边或者弧表示关系,从而通过图来表示作者的知识。目前知识图的重点已经转向一般的知识表示形式[4]。

本文中,我们挑选了一些具有一定代表性的汉语中具有"CAU"关系的操作符,并且根据与这些操作符关联的不同的操作数类型对这些挑选出的操作符进行分类。主要目的是通过这些具有"CAU"关系的操作符建立文本知识的自动抽取。

二、"CAU"词在知识图中的表示

在知识图理论中,"CAU"语句,像"IF - THEN"这样的语句,是通过"CAU"操作符和与其关联的操作数来表示的。也就是说,"if A then B"(或者"A 是 B 的起因")是用如图 1 所示的结构来表示的。这里,"CAU"是起因操作符,A 和 B 是相应的操作数。"A CAU B"的意思是:A 的出现或变化可以引起 B 的出现或变化。

这样,任何具有起因意义的短语(句子)就可以用一个起因操作符和相应的操作数来表示。但"IF - THEN"这样的语句并不总是等价于"CAU"语句。文献[1]中讨论了知识图中"IF - THEN"语句的框架表示以及等价的逻辑表示。

如果 *A* 的增量 △*A* 引起 *B* 的增量 △*B*,则我们可以说"如果 △*A* 出现则 △*B* 出现"。因此,"△*A* 出现"与"△*B* 出现"之间存在起因关系。按这种解释,"IF - THEN"关系就与"CAU"关系联系在一起。汉语中,有很多单词和词组是具有"起因"意义的,同时,按照与这些具有"起因"意义的操作符相连接的操作数的不同类型,我们可以给这些操作符分类。

例:考虑句子"按照计划我们写文章"。句子的主要部分的表示如图 2 所示。这里,中心词是"计划"和"写","计划"是"写"的原因。"按照"是表示"起因"意义的操作符,它连接两个操作数:"计划"和"写"。

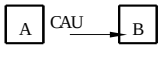


图 1　起因关系

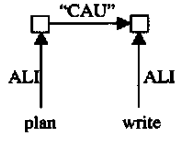


图 2　部分句子图

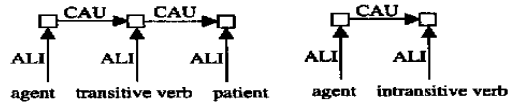


图 3　句子的表示

知识图中我们使用图 3 所示的结构表示句子的主要成份。图 2 中虽然我们使用了标号"CAU",但是很明显"计划"和"写"之间不存在施受关系。我们之所以仍然使用标号"CAU"的原因是这样的:"写"的"内容"或"方式"与"计划"之间存在直接的因果关系,只是"内容"或"方式"在句子中没有表示出来。因此,通过一定的概念引申,"按照"可以看作是"起因"操作符。虽然使用一个简单的表示起因("CAU")关系的弧就可以表示它,但是它的标号应该看成是经过扩充的起因关系,可用小写字符"cau"来表示。(但为了简单,本文中没有着意区分"CAU"和"cau"的差别。)这种现象在自然语言中非常普遍。

三、挑选的“CAU”操作符

就像英语中一样,一些汉语中的词是非常明显的“起因”单词。英语中具有“起因”关系的操作符,像“causes”和“leads to”,是很容易辨别[3]。然而,在例子中我们已经看到,操作符可能不是很明显地表现出“起因”内容,仅仅当我们对单词的意义进行更详细的分析后(词图正是完成这项工作),“起因”内容才会变得清楚。

如果A和B按某种方式相联系,在知识图理论中这种联系是通过包含有A和B的知识图来表示的。知识图的结构表示了A和B的联系方式。一个“起因”关系只是这个图中的一部分,就像例1中我们对“按照”的分析那样。然而,用来表示“计划”影响“写”的详细的图比仅用一个“cau”关系来连结“计划”和“写”的图要复杂的多。这个复杂图会因为引申图中的概念而变得很大很大,从而得到表示关系的词的很多很多意义。

知识图中强调“结构就是意义”。每个词应该有一个刻划这个词的词图,并且这个词图应该与其它词的词图有差别。因此,原则上讲,我们有责任来研究每个这样的具有“起因”意义的词,并且给出一个精确刻划的词图。然而,由于篇幅所限,文中没有给出每个“CAU”操作符的详细词图,因此,可能会有这样的情况,同样的图用来表示两个不同的词,但是,如果我们将注意力放在节点上的话,不同的含义就由这些节点所表示的内容来体现。

从文本中抽取“起因”关系将会获得一个“CAU”关系类型的弧的集合,这个集合可以用来建立有向图。这个有向图又可以用在专家系统或决策支持系统中。在很多情况下,图中的节点是由名词来标号的。为建立专家系统,应将注意力放在具有“起因”关系的图上,因此,只需表示具有起因意义的操作符和由这些操作符所联系的操作数。

句子“IF p THEN q”可以用p—CAU→q来表示。为了与“CAU”关系的解释相一致,我们选择“IF p THEN q”意味着“p的出现(真值)引起q的出现(真值)”。在纯粹逻辑中,p和q之间可以没有任何关系。然而,在我们建立专家系统时,重要的是详细地确定p和q在什么情况下具有起因关系。在“IF p THEN q”中,p可以是q出现的条件。当q描述的是一个过程、一个化学反应时,p可以表示数值条件。例如,如果p表示“温度”,q表示“分子A和B组合成分子C”,我们可以在“温度”和“组合”之间用起因关系来表示。这个例子说明起因操作符可以连结一个名词和一个动词作为其操作数。由于在知识图理论中,原则上讲,“CAU”关系一般连结的都是名词,所以,这些动词都可以用名词来替换(例如,可以用名词“combining”或“combination”来替换动词“combine”)。下面我们通过一些例子来分析所选的汉语“起因”词。这些词是从文献[8]中得到的。同时,给出所选例子中的操作数,并设法挑选出其中与“CAU”操作符关联的名词项。

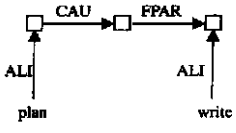


图4 “计划”与“写”之间的关系

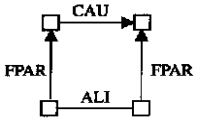


图5 “按照”的表示

第一组:名词与动词相关联

1. 操作符:按照

例:按照计划,我们写文章。

分析:“写”的“内容”或“方式”受“计划”的影响。由于“写”是一个复杂的框架,“写”的“内容”或“方式”是通过“写”的“FPAR”关系来确定,即“计划”相当于“写”的条件,知识图表示如图4。

“写”的“内容”或“方式”应该显示出与“计划”的一致性,这可以通过ALI-关系来表示。因此,我们可以用图5来表示这种“起因”和“一致”。

名词操作数:“计划”、“文章”。

解释:不同的计划导致不同的文章,因此“计划”与“文章”之间存在“CAU”关系。

表示:计划 —ALI→□—CAU→□←ALI—文章

像这样的“起因”操作符还有“根据”、“凭”等。

第二组:名词与名词相关联

2. 操作符:来源于

例:知识来源于实践。

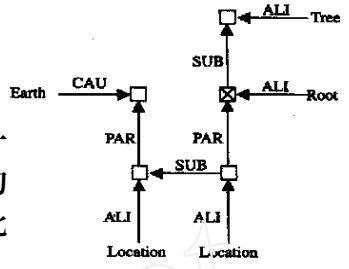


图 6 “tree roots in earth”的表示

分析:这里使用了比喻。正像土地对树木生长是必须的一样,实践是获得知识所必须的。作为条件,我们使用一个简单的“CAU”关系来表示它。让我们来较深入地研究语言中的这种比喻。对于“tree roots in earth”我们可以构造如图 6 所示的知识图。这个图仅仅描述了树的根与土地之间的几何关系。当我们说“knowledge roots in practice”,只需要将其中的词替换掉即可。然而,这句话的目的是表达“实践”导致“知识”。为了证明词“根”的使用,在几何关系之外应该辨别起因关系。树的生长依赖于,或起因于,土地的营养、光以及水份等。树木的根扮演着这个复杂过程的中间角色。土地这个框架概念含有“养分”这个概念作为它的一部分:

Nutrients —ALI→□—FPAR→□←ALI—earth

并且,概念“养分”与概念“树木”之间存在起因关系。通过类推,“实践”应该包含有“知识”生长的起因。众所周知,思维就是某些东西的联系。在知识图理论中,我们可以说明“实践”的某些方面与“养分”类似。“实践”伴随着精神过程,“思维”是知识增长的原因,也就是思维图(mind graph)增长的原因。“养分”包含在这个过程中。

这个比喻,正像所有的比喻那样,是非常复杂的。我们详细分析这个例子,因为它给出了知识图理论核心的一个反映。对于起因操作符“来源于”来说,其复杂的表示可以用图 7 来说明。

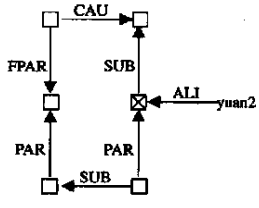


图 7 “root in”的表示

我们的分析得到下面的结论:起因过程是从“实践”的框架结构中的某些部分到“知识”。我们选择使用一个简单的 CAU 关系来表示它。

名词操作数:“知识”、“实践”

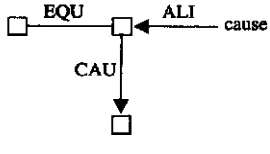
解释:正如我们已经分析的那样,通过使用比喻,隐含着一个实际的起因关系。

表示:没有抽取出与名词操作数相关联的“CAU”连结。

3. 操作符:在于

例:事物发展的根本原因在于其内部矛盾。

分析:这里,很明显指出了发生的原因,因此可以表示为右图:



名词操作数:“矛盾”、“发展”

解释:这里,操作数具有起因关系很明显,我们可以非常清楚地看到与前面几个例子的区别。但是,寻找具有起因关系的精确的操作数是非常困难的。“矛盾”与“发展”之间的实际关系存在模糊之处,甚至当对概念“发展”作进一步扩展之后也同样。

表示:矛盾 —ALI→□—CAU→□←ALI—发展

到这里,我们发现所给的例子中并不是都能挑出名词“CAU”连结——即没有找到期望的名词操作数,而正是这些名词“CAU”连结在建立专家系统时才非常有用。虽然我们已经确定了一些具有起因意义的名词对,但是同时我们也发现一个事实,虽然操作符确实是“CAU”操作符(或“cau”操作符),但是在很多情况下不包含直接的名词操作数,即基本的起因关系没有明确包含,仅仅是通过一定的联想——概念意义的扩充——我们才可以发现为什么句子中存在起因操作符。

我们现在面对的是表示法的确定问题。对所有的“CAU”操作符,或者我们给出详细的不同的词图,这样有助于确定对应的操作数,或者我们采取简单的“CAU”连结来表示这些复杂图。虽然在知道详细词图的情况下,“CAU”操作符所关联的基本因果关系也不是可以直接得到的,因此我们选择的是第二种方案。我们已经看到,对于这些操作数,常常可以确定名词对,但这些名词对之间没有直接的因果关系。当然,这样抽取出来的结果的质量有些粗糙。

为了从科技文本中抽取知识,表示需要简单化。即使实际的“CAU”操作符非常复杂,但是实际使用上都用的是其简单的形式。通常使用的“CAU”操作符都限定在一个非常小的集合上。情况4和5中所表示的都是“摩擦”和“热”之间的起因关系,操作数很容易确定。结果都可表示为:摩擦—ALI→□—CAU→□←ALI—热

4. 操作符:产生

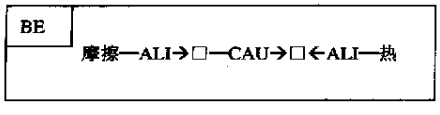
例:摩擦产生热。

分析:起因过程很明显,因此,我们用一个基本的“CAU”关系来表示。

5. 操作符:产生于

例:热产生于摩擦。

分析:与4的不同之处多加了一个“于”,从而改变了因果关系。我们知道,主语是一句话所要表达的话题(topic),因此,4与5中的操作符的不同之处是所引出的话题不同。4中的话题是“摩擦”,而5中的话题是“热”,5中的例子可以看作是4中例子的被动(逆)表示。如果忽略这一点,那么其基本含义是一样的。如果要区分它们的细微差别,则这里的例子可表示为下图。



像这样的“起因”操作符还有“安含着”、“导致”、“致使”等。

通过这些例子,我们看到“CAU”操作符使用了很多比喻。正如在第一组所遇到的问题那样,经常使用比喻会造成很大的问题。

第三组:句子与句子相关联

这一组包含逻辑词,我们在第一节中已经论述过它们用“CAU”操作符的表示。参考文献[1]专门研究过逻辑词。

6. 操作符:如果

例:如果天气晴朗,我们可以出游。

分析:这句话中不存在直接到起因关系。天气状态并不影响出游,而影响的是做出出游的决定。但是,这句话中又没有提到做决定。省略在语言中经常碰到,这是确定“CAU”操作符的操作数的主要困难之一。

名词操作数:“天气”、“出游”

解释:逻辑语句。

表示:没有抽取出与名词操作数相关联的“CAU”连结。

7. 操作符:所以,因此

例:他学习努力,因此(所以)他取得很大进步。

分析:“所以(因此)”直接涉及的起因是“学习”,因此我们使用了“CAU”操作符。但是确定操作数需要详细讨论。句子描述的是一个过程,这个过程本身可以看作是操作数。简单的讲,可以选择“学习”和“进步”来作为操作数。

进一步分析需要对“学习”和“进步”进行扩展。如果“学习”意味着“知识的增加”,“进步”意味着“某些东西增加”的话,对分析这句话给了我们一个新的启示:我们可以用“如果知识增加,则某些东西增加”来表示这句话,这样又归入逻辑词的范畴。

名词操作数:“学习”、“进步”

解释:“学习”=“知识的增加”,意味着“某些东西的增加”=“进步”。

表示:没有抽取出与名词操作数相关联的“CAU”连结。

注释:对这一组“CAU”操作符的一个很重要的注释是,这些操作符在句子中“可能”,但不“总是”,表示起因关系。

像这样的“起因”操作符还有“从而”、“既然”等。

第四组:名词与句子相关联

8. 操作符:使得

例:大雪使得交通瘫痪。

分析:这里,“CAU”关系直接从“雪”指向“交通”。

名词操作数:“雪”、“交通”

解释:这是一个非常普通的起因关系。

表示:雪 —ALI→□—CAU →□←ALI—交通

注释:这一组中包含有一些没有提及的名词,但是这些没有提到的概念却确实包含有“CAU”关系。

像这样的“起因”操作符还有“鉴于”、“凭借”、“由于”等。

第五组:其它

9. 操作符:为了

例:为了方便使用计算机,我们研究自然语言处理。

分析:使用计算机是目的,而这个目的是通过研究来达到。“为了”本身不是一个“CAU”操作符,不表起因关系,但是,“为了”却表示“目的”,而“目的”含有起因关系。

名词操作数:“研究”、“使用”

解释:“为了”表示的是目的,不是“起因”。

表示:没有抽取出与名词操作数相关联的“CAU”连结。

像这样的“起因”操作符还有“引起”等。

虽然这里所列举的所有操作符均是“起因”操作符,或者从中可以导出“起因”操作符,然而,在日常生活中,在非技术领域内的交流要求听者和说者具有一定的解释能力。我们这里所列举的例子中,严格地讲,很多情况没法抽取出相应的、我们希望的“CAU”连结,虽然在对句子图进行扩充分析后确实具有“CAU”连结。在这种连结中,有趣的是含有“起因”的“CAU”操作符,在分析中我们发现,“原因”是某种意义上的起因,但是常常是由人们所做的决定而引起的,但是词“决定”常常是没有提到的,因为它已经隐含在词“原因”中。

这里我们已经给出了很多具有起因意义的操作符,由于人们解释一个句子时要用到背景知识,这给自动抽取"CAU"操作符及其所关联的名词操作数造成了巨大的困难。然而,技术文献与日常生活所使用的语言有差异,需要的背景知识非常少。目前我们正在研究从科技文本中抽取"CAU"关系的实例。

四、结论

这里所收集到的汉语"CAU"操作符显示出一些性质,这些性质对于通过文本建立专家系统中的知识自动抽取是非常重要的。

第一个性质是逻辑含义与起因关系之间的差异。像"如果 2 ×2 = 4,则北京是一个大城市"或者"如果 2 ×2 = 5,则北京是一个小城市"这样的句子,其中不存在任何起因关系。从逻辑的观点来看却都含有逻辑关系,但是很显然,没有人会从这样的句子中抽取"CAU"连结。

然而,在自然语言中,用词并不总是恰当的。让我们考虑"摩擦"和"热"。从"摩擦引起热"中立即可以建立"CAU"连结。"如果摩擦出现,则热出现"是一个完全的逻辑结构。这个陈述隐含着起因关系,但是当我们把"如果"、"仅当"和"如果...则"这样的操作符从"起因"操作集合中去掉的话,就会漏掉这样的起因关系。因此,表示"CAU"关系的操作符强迫我们要考虑所关联的命题"摩擦出现"和"热出现"。当"摩擦"和"发热"都可以看成是一个过程时,则它们之间确实存在"CAU"连结。因此结论是:在处理句子与句子之间的操作符时要格外细心。如果存在"CAU"连结,应该有"出现"或"变化"这样的词存在,同时,连结的仅是名词与名词。

第二个性质是由明确提到"原因"或"为了"这样的词用来表达目的的一些操作符。包含这些"CAU"操作符的句子的注意力放在人的推理过程上。不过,在考虑这些操作符的操作数的层次上,可能存在起因关系,这些起因关系是通过目的的方式来表示。正如对"IF-THEN"句子那样,其中的起因关系可以通过考察是否有名词变化和出现,来确定具有相同意义的、能够辨别起因关系的标准句。

第三个性质是语言中比喻的使用。用"来源于"这样的操作符连结的句子通常都非常复杂。这种应用有一部分来源于汉字构词法——表达了一些词水平以下的关系;另一部分是由于语言中比喻的使用,由于使用了比喻,说话者没有也没有必要对他所讲的话做进一步的解释,相信听者可以从中推断出含糊不清的话语中的意义。

对于我们所要的抽出知识到过程,它可以帮助我们寻找关联的名词操作数,并且同时获得具有相同意义(大致相同)的标准句。

在所有这三种性质中有一个共同点:确定有"变化"或"出现"的名词操作数,并且或多或少的保留原来句子的意义。

感谢荷兰 Twente 大学教授、西北工业大学顾问教授 C. Hoede 的帮助。

参 考 文 献

[1] Hoede, C. And L. Zhang, *Word Graphs: The Third Set*, in: Conceptual Structures: Broadening the Base (H. S. Delugach and G. Stumme, eds.), Proceedings of the 9th International Conference on Conceptual Structures, Stanford University, California, U. S. A. Springer Lecture Notes in Artificial Intelligence, Vol. 2120, ISBN 3 - 540 - 42344 - 3, 2001, 15 - 28

[2] James, P., *Knowledge Graphs*, in Linguistic Instruments in Knowledge Engineering (R. P. Van de Riet and R. A. Meersman, eds.) ISBN 0 - 444 - 88394 - 0, 1992, 97 - 117

[3] Vries, P. H. de, *Representation of Scientific Texts in Knowledge Graphs*, Ph. D. Thesis, Rijksuniversiteit Gronin-

gen ,Groningen ,The Netherlands ,1989

[4] Willems ,M. , *Chemistry of language : a graph theoretical study of linguistic semantics* , Ph. D. Thesis ,University of Twente ,Enschede ,The Netherlands ,ISBN 90 - 9005672 - 9 ,1993

[5] Whitehead ,A. N. and B. Russell , *Principia Mathematica* , Cambridge University Press ,Cambridge ,1925 ,2nd edition

[6] Huang ,K. ,Introduction to Expert Systems ,Southeastern University Press ,1988

[7] 刘小冬 ,李学良 ,张蕾. 知识图综述. 工程数学学报 ,2000(17)增 :33 - 40

[8] 新汉英词典. 北京 :外语教学与研究出版社 ,1988

(上接第 37 页)

[3] Voorhees E M ,Harman D K ,The sixth Test REtrieval Conference (TREC-6) , Gaithersburg ,NIST ,1998

[4] Salton G ,The SMART retrieval system-experiments in automatic document processing ,Prentice Hall ,1971 :115 - 411

[5] http : //morph. ldc. upenn. edu/ Projects/ Chinese

[6] Gao J F ,Nie J Y ,Zhang J ,et al ,Improving query translation for CLIR using statistical models ,ACM SIGIR '01 , New Orleans ,2001 :96 - 104

[7] David Hull ,Using statistical testing in the evaluation of retrieval performance , In Proc. of the 16th ACM/ SIGIR Conference ,1993 :329 - 338

会议消息

中国中文信息学会与国际中文计算机学会于 2003 年 8 月 3 日 —6 日在沈阳市召开“第二十届东方语言计算机处理国际学术会议”,会议由东北大学承办。

2003 年 8 月 8 日 —11 日中国中文信息学会与兄弟学会在哈尔滨市召开“全国第六届计算语言学联合学术会议(JSCL —2003)”。会议由哈尔滨工业大学承办。

上述两次会议的征文内容如下:

1) 计算语言学的理论研究;

2) 汉语的词汇、句法和语义;

3) 语料库建设、语料加工技术及基于语料库的语言分析技术;

4) 汉语的文本分析与生成;

5) 机器翻译系统、技术及评测方法;

6) 文本智能检索、文本自动分类、文本过滤及自动文摘;

7) 汉语语音识别与语音合成;

8) 智能型汉字输入方法;

9) 其它。

征文截稿日期等重要信息另行通知。