

歧义、系统歧义和语境

钱 树 人

(南京大学计算机科学系)

【摘要】 本文对歧义现象,特别是对语言片段的歧义理解进行了剖析,并研究了不同语境对理解歧义的影响。进而提出了默认语境,系统语境和系统歧义等概念,并简要地介绍了汉语语言片段歧义分析模型系统 CAAMS。

一、引 言

自然语言中的歧义现象十分普遍和复杂。歧义现象对于计算机处理来说是不希望碰到的,总是想方设法在描述或实现中加以消除或排除。例如,在编译中常常会有“向前偷看”这类操作,就是为了消除分析源程序的歧义。IF 语句在早期的 ALGOL60 版本中是具有歧义的,以后采用在语法描述上加语句闭括号或在语义描述上对其子成分加上限制约束以排除其歧义理解,以及在编译实现时按某种方式理解强行消除其歧义等方法。在对自然语言的分析理解中,很快发现其歧义现象远为复杂得多。

在自然语言理解中具有“天然歧义”现象,即人们有意识地运用“歧义”以达到语言表达的特定效果,例如“双关语”和比喻等。这种使用方式是歧义现象的积极方面,但和目前的自然语言计算机处理的水平相差甚远。更多地则是无意中引进了歧义现象,从而产生一些不希望的误解和含混。

歧义分析是自然语言理解的一个重要方面。自然语言本身,即使是受限的自然语言,也远比程序设计语言和形式语言复杂,因此其歧义的理解和分析自然更为困难和复杂。

自然语言的理解主要分真值理解和意义理解两方面。真值理解主要是对自然语言句所叙述的事实、事件和行为等的真值性进行判断,而意义理解则是对自然语言片段所反映的信息和特征、属性进行相应逼近需求的提取,以及建立相应的表示。

这两类不同的理解方式中当然均存在歧义问题。歧义分析理解具有不同的层次,和语境,语法语义语用,系统,用户需求等等有着密切的联系。本文主要从计算机处理角度而不是从语言学角度来讨论它们。

本文1992年9月9日收到。

二、语境对自然语言真值分析理解的影响

文本自然语言（即书写用的自然语言）的语境主要指所要理解的语言片段的上下文及其限制和约定。所以，有的使用 context（一般译为上下文）来表称语境，但它和程序设计语言，形式语言中所指的“上下文”仍有所不同，宜另用一个术语（如 Environment）来表称也许更好一些。

程序设计语言或形式语言中的上下文实际上是指“紧上下文”，形如 XwY 。其中， w 为所要导出的（分析理解）成分，而 X, Y 为紧靠 w 的上文和下文。文本自然语言和话语自然语言中所述的上下文是广义的，不一定是紧靠的，还可能是隔句的，而且主要是上文。

例 1. “他上午八时准时走进机房后，开启电源，使计算机开始工作，休息一会儿，泡茶，喝茶和注视控制屏幕上的一切变化。”

这是文本自然语言，描述了一连串事件，有时间有地点有时序有动作。

显见，“注视控制屏幕上的一切变化”是一个缺省了主语“他”的陈述子句，对其理解可以仅考虑其真值性也可以进行意义、结构、信息的多角度分析。即使仅考虑其真值性也是相当复杂的，其语境由相应的一连串的语境条件组成，这些语境条件有的对其理解具有约束作用，有的并无明显约束作用。

例 1 所述的句子可从许多角度理解。

从时序上看，这一连串事件基本上是有顺序的，而且不应当颠倒。但“喝茶”和“注视控制屏幕上的一切变化”可以是串行的也可以是平行的，而且串行和平行均不会产生矛盾。

从前提上看，也呈现着复杂的情况。以“注视控制屏幕上的一切变化”作为理解点，则其前面的一连串事件应视为一连串的语境条件。在理解点处存在的所有语境条件构成“理解点语境”（简称为“语境”），由此可看出，语境具有动态性逼近性和时序。“休息一会儿”，“泡茶”，“喝茶”并不是“注视控制屏上的一切变化”的必要前提，但它们是语境条件，故可称为“理解点无关语境条件”。而“开启电源”和“计算机开始工作”则是“注视控制屏幕上的一切变化”的“理解点有关语境条件”，而且是“理解点有关的相容语境条件”。

例 2. “时至中午下班，他关掉电源，将门窗关好，注视控制屏幕上的一切变化”。

在例 2 中，“关掉电源”是“注视控制屏幕上的一切变化”的“理解点有关语境条件”，但它是“理解点有关的不相容语境条件”。因为，“关掉电源”的条件下再去“注视控制屏幕上的一切变化”是不符合人的认知心理的。由此可以推知，从真值理解角度看，例 1 中的“注视控制屏幕上的一切变化”具有真值，而例 2 中的“注视控制屏幕上的一切变化”则具有假值。

再进一步分析，在例 1 中没有“开启电源”这一子句，而例 2 中也没有“关掉电源”这一子句的情况下，如何理解“注意控制屏幕上的一切变化”的真值性呢？为此引进概念“理解点的默认语境条件”且作为“理解点有关的相容语境条件”处理。亦即，分别默认“电源开”或“电源关”这两个语境条件存在，从而“注视控制屏幕上的一切变化”的真值理解结果为真。

由上述讨论可以看出语境条件的约束对自然语言的真值理解也是很重要的。为此引入相应的概念描述如下。

(1) 存在所要真值理解的事件或命题, 记为 PS 。

它是一个叙述得当的句子或子句。换言之讲, 如果不考虑语境等约束, 该句子的表达是完整的、合理合法的。亦可讲, 在默认语境下, PS 的真值理解结果为真。

(2) 对 PS 进行真值理解时, 存在一个语境 $EN(PS)$, 它是一个语境条件集, 即

$$EN(PS) = \{P_1, P_2, \dots, P_n\} \quad (n \text{ 为有限})$$

这里的 P_1, P_2, \dots, P_n 为非默认语境条件(或称显式语境条件)且可以规定所有的 P_i 与 PS 的理解有关及 P_1, \dots, P_n 彼此之间无矛盾。

当 $EN(PS)$ 为空集, 则化归为简单真值理解。即, 所有的语境条件均是默认且相容的, 而且 PS 本身是成立的, 故其真值理解结果为真。

当 $EN(PS)$ 中存在一语境条件与 PS 不相容时, 则 PS 的真值理解结果为假。

当 $EN(PS)$ 中所有的语境条件均与 PS 相容(或无关)时, 则 PS 的真值理解结果为真。

(3) 设 PS 的后继理解事件为 SPS , 那么 SPS 的语境可以由下法构成:

$$EN(SPS) = \begin{cases} EN(PS) \cup \{PS\} & \text{当 } PS \text{ 的理解为真时} \\ EN(PS) & \text{当 } PS \text{ 的理解为假时} \end{cases}$$

当然上述讨论仅涉及到语境作用的很小部分。至于话语自然语言(即对话用自然语言)的语境问题则比文本自然语言更为复杂。其原因在于: 话语自然语言中允许语言成分更多的省略; 语言的表达更为自由和多样化; 具有更多的非规范特征; 其语义不但和当前所述和先前所述的语句有关, 而且还会依赖于口吻、语调、表情、姿势、动作以及说话时的场景等等。例如, “你, 你, 你和你们的班长一起到办公室来开会!” 这句话从文本自然语言来看是不好理解的, 但在特定的对话场景中是可以理解的。因为这时采用手势来指明“你、你、你”, 而不是采用上下文来指明的。

语境对自然语言意义分析理解的影响则更为复杂得多, 有待进一步研究和另行阐述。

三、语境对歧义理解分析的影响

在自然语言的分析理解处理中, 歧义的含义、刻划、分析、判定和排除是十分重要的。

歧义的含义首先与理解的要求有关。真值理解和意义理解下的歧义性具有不同的含义。如果在 $EN(PS)$ 语境下尚不能判定所述命题的真假, 则称在 $EN(PS)$ 语境下命题 PS 具有歧义性, 更确切地称为具有真值理解歧义性。由于语境和语境条件的确定和提取是会随着系统的不同而有所不同, 故应进一步引进在真值理解下的系统歧义性。

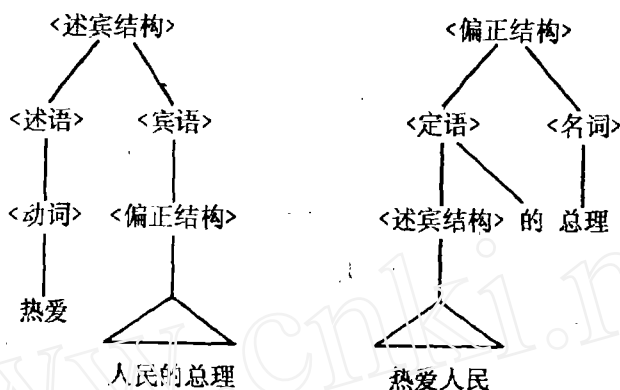
通过对语言片段(或称语段, 它可以是词、词组、子句、句子乃至一段文本)的歧义现象的分析, 可以得到许多有益的启示, 从而进一步完善语言学的各种体系。对特定的应用系统而言, 涉及的自然语言可以加以合理适当的限制, 使之歧义性减少, 使应用系统既符合实际使用要求又更为有效。但“减少”并不能保证完全排除歧义现象, 因此应要求系统能自动进行歧义分析。为此, 正确理解歧义语段就必须借助语言学等方面的成果, 提取相应的知识, 并使之形式化和工程化。

意义理解的歧义更为复杂, 为此对其歧义性引入下列概念。

定义1. 一个语言片段为歧义的当且仅当该语言片段具有两个或两个以上不同的意义。

“什么是意义?”这一概念是有不同认识的,如理性意义,内涵意义、社会意义,情感意义,反映意义,搭配意义等等。由此可见,判断语言片段是否歧义,获取语言片段的意义将是必须的。获取和分析语言片段的意义将至少要考虑下列几个方面:意义的表示方法和逼近程度;分析组成语言片段的各个词的词汇意义;分析词汇意义之间的组合搭配关系;分析语言片段的结构意义;等等。换言之,分析语言片段的歧义主要来自词汇的多义现象,词汇之间的搭配有多种形式或多种意义,语言片段的结构分析具有多种不同的选择等原因而造成的。

例3 “热爱人民的总理”这一语言片段从结构分析看有两种形式。



例3 (a)

例3 (b)

这个例子中的各个词汇的意义和词性均无歧义,但语法结构有歧义,从而导致意义理解歧义。但当存在某种语境时则可能消除其歧义。例如,

“全国老百姓热爱人民的总理。”

那么,“全国老百姓”是语境部分,而“热爱人民的总理”只能按(a)理解,且排除(b)的理解。

“追忆热爱人民的总理。”

那么,“追忆”即是语境部分,而“热爱人民的总理”只能按(b)进行意义理解。

例4 “老张和老李的妻子”这一语言片段从结构分析看也有两种形式。

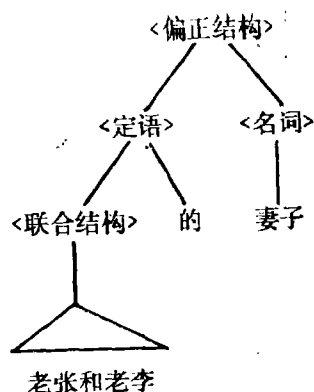
如果知道老张和老李的属性、特征和背景知识(应视作隐式语境),也可能排除其歧义理解。

当老李是女的,则该语言片段的意义理解结果为该语言片段有误(语义错误)。

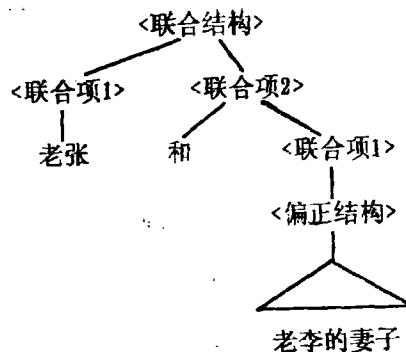
当老张是女的,则应理解结果为例4(b)。

当老张和老李均为男的,则该语言片段的意义理解既可以是例4(a)或例4(b),即存在歧义理解。

再如“老张已年老且已生前列腺炎,老张和老李的妻子到医院去探视。”仅从排除“老张和老李的妻子”的歧义理解,根据“老张已生前列腺炎”即可推出“老张是男的”,故其意义理解应当如例4(a)。这里还隐含着许多信息,如老张病较重已住医院,老张的妻子和老李



例4 (a)



例4 (b)

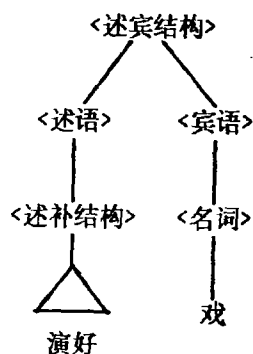
的妻子关系相当好等等,但在这里是次要的且可忽略的。由此可见,有用信息的提取是具有逼近性和层次性的。

例5 “演好戏”这个语言片段由于“好”有两种词性,可以用作付词和形容词,虽其词义相近,但由于组合顺序不同,使该语言片段有不同的结构划分。

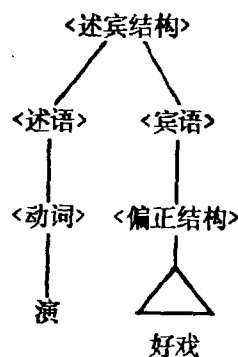
例如“张三演好戏!”在“张三演戏态度不认真”的语境条件下,其含义的理解应取例5(a)。在“张三平时演的戏质量差”的语境条件下,其含义的理解应取例5(b)。

四、系统歧义和模型系统 CAAMS

从上面的分析中看出语境和语境条件对理解和歧义分析有很大的影响,语境条件的提取有很大的难度,通常与系统实际需求有关。例如,词典的大小,词条的定义和解释,句型的



例5 (a)



例5 (b)

多少及其对应表示等等在实际系统中总是有限的,而且由于实用性和有效性等因素常常需要对自然语言的覆盖面有一定的要求。为此引进了若干定义来刻画这种现象。

定义2 语言片段在系统中的某种表示称为该语言片的系统表示。这种表示可以是树、

谓词、语义网、格、等等，而且可以引进映射：

$$NLT: L \rightarrow 'LS$$

L 为自然语言句集，LS 为系统表示集，映射 NLT 实现其变换。这种变换可能是近似的，仅提取必要信息而不是完全信息，从而可能出现多对一等复杂情况。

定义 3 若语言片段存在一个系统表示，则称该语言片段是系统合法的，或称为(系统)有意义的。否则称该语言片段是系统非法的，无意义的。

定义 4 若语言片段存在且仅存在一个系统表示，则称该语言片段是系统非歧义的。而若存在两个或两个以上的系统表示则称该语言片段是系统歧义的。

定义 5 若语言片段存在 n 个系统表示，则称该语言片段是系统 n 义的。能给出所有语言片段的全部系统表示的功能称为“系统全解”功能。

上述定义中的语言片段的合法和排法，歧义和非歧义均冠以“系统”，其原因在于理解依赖于实现系统中所采用的“系统表示”和所具有的代表能力。由于自然语言理解的复杂、困难和实际需要，建立“有限覆盖”自然语言和“逼近提取”有效信息是合理的，从而有利于严格的刻划和评测以及形式化研究，有利于计算机处理及满足用户的需求。

CAAMS 系统是一个针对汉语语段进行歧义分析具有系统全解功能的模型系统，简称为“汉语语段歧义分析模型至统”。

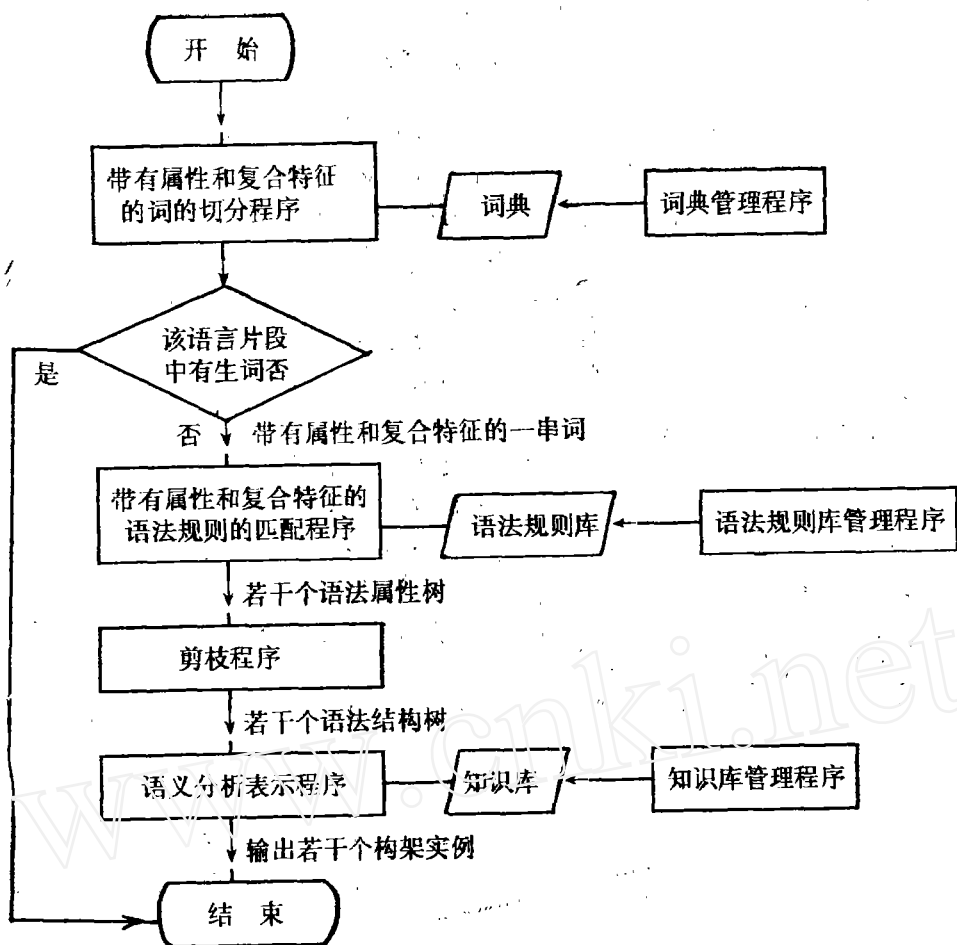
CAAMS 以汉语的语言片段为分析理解对象，在隐式语境和理性意义理解的前提下实现的。

CAAMS 的流程概图如下。

CAAMS 系统得益于一种很有用的描述工具——属性文法，定义了带有属性特征的词典和语法规则库，并建立了相应的算法。系统能借助词典、语法规则库以及有关算法对语言片段进行分析，给出所有可能的语法结构树，即具有“系统全解”功能。当无语法结构树产生则表示该语言片段为系统非法；当仅有一个语法结构树产生则表示该语言片段为系统合法且系统非歧义的；当有两棵或两棵以上的语法结构树产生则表示该语言片段为系统合法且系统歧义的。CAAMS 中定义了带有属性特征的词典描述语言 DDL 和语法规则描述语言 SRDL，并提出了一系列算法。在 DDL 中，一个词的定义包括两部分：名和相关的属性。属性可由若干个不同的属性元组成。系统将语言片段切分以后，生成一串带有属性的词。在 SRDL 中，允许为一个规则定义若干个属性赋值语句和条件。属性赋值语句实现了属性在规则间的传递，条件则定义了规则的适用范围。CAAMS 借助语法规则库和带有属性的语法匹配算法，生成若干个语法属性树。且借助属性计算算法对若干个语法属性树进行剪枝，生成零个、一个或多个语法结构树。

CAAMS 具有可扩充性。它可以增删改词典中的词、词义或属性，从而可使系统“认识”的词增加或更精确，使词典不断地扩大。它可以增删改语法规则库，从而允许系统不断地扩充“认识”和“适用”的语法规则。这种改动无疑会影响“系统歧义性”的分析结果。当然，改动词典和语法规则库需要仔细地处理以免引进“不一致”和混乱。由于整个系统的分析理解算法是不动的，故正确扩充词典或语法规则库后，将会增加“系统合法”的语言片段，其“系统歧义”的分析结果也愈加“逼近”实际。

在 CAAMS 中采用的系统表示是语法结构树。由此自然可以推出判断语言片段合法与



否，歧义与否的主要依据是由 CAAMS 分析理解该语言片段所得的语法结构树 来决定的。其规定由定义 2、定义 3 指明。

语义分析程序完成对前面所生成的语法结构树的语义解释，在该系统中即是 将语法结构树映射成其意义的表示。各种自然语言理解系统由于要求不同方法不同，语义解释的逼近程度（有效信息的提取，表述的深度和广度）不同，语义的表述形式也是多种多样的。在 CAAMS 中，采用基于构架的形式来表示，即一语言片段的语义用该语言片段主词的构架实例来表达，而且其构架实例的形式用 BNF 形式描述如下。

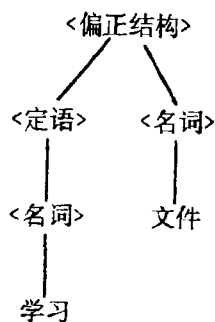
〈构架实例〉::=〈实例名〉“[”〈体〉“]”

〈体〉::= |〈属性名〉“=”〈属性值〉|〈体〉; 〈体〉

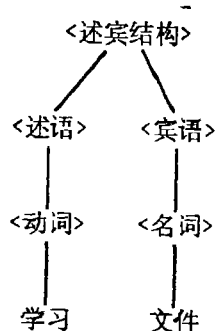
〈属性值〉::=〈词〉|〈实例名〉|〈属性值〉“|”〈属性值〉

其中：实例名用于标识各个实例；“|”号用于表示各个值之间的地位是相互平等的，即一个属性具有多个值；属性名即相应实例名中的词在词典中所具有的。

例 6 “学习文件”在 CAAMS 系统中能求得其全解，即可构造出所有的两棵语法结构树：



例6 (a)



例6 (b)

这两种语法结构树均符合我们的习惯，而且反映了相应的语义。经过语义分析和映射，分别得到构架实例如下：

文件[定语 = 学习] 对应例6(a)

学习[宾语 = 文件] 对应例6(b)

“学习”，“文件”分别是实例名，也是语言片段语义所对应的构架的主词。主词和语法结构有关，例如主谓结构以谓语动词为主词，述宾结构和述补结构以述语动词为主词，偏正结构以中心词为主等等。

Language Environment and Systematic Ambiguity

Qian Shuren

(Dept. of Computer Science, Univ. of Nanjing)

Abstract

This paper analyzes an understanding of language segment ambiguity in detail, and effect on ambiguity in different language environment. Furthermore, some conception such as default language environment, systematic language environment, Language environment condition, systematic ambiguity, are specified. And the CAAMS, Chinese segment Ambiguity Analysis Model System, is intrucluced,

参 考 文 献

- 1 Graeme Hirst, Semantic Interpretation and the Resolution of Ambiguity, cambridge univ. press. 1987.
- 2 徐烈炯, 语义学, 语文出版社, 1990.
- 3 F. Guenther and S.J. Schmidt edit, Formal Semantics and Pragmatics for Natural Languages, D. Reidel Pub. co. 1979.
- 4 M.D. Harris, Introduction to Natural Language Processing, 1985.
- 5 刘开瑛, 郭炳炎, 自然语言处理, 科学出版社, 1991.

www.cnki.net