第4章　思维运作

第10讲 语言理解

理解力是心智能力的重要方面。一般理解力包括时间理解力、空间理解力和因果理解力，强调对概念及其关系的把握。

在人类的心智活动中，最能全面反映理解力表现的莫过于是对语言意义的理解。

当然，要使机器具备语言理解能力并非是一件轻而易举的事情，这其中会涉及到几乎人类心智和文化的所有方面。

本节我们专门讨论语言理解方面的基本内容，特别是有关汉语机器理解的研究工作和面临的困难。

1、多尺度意群分割

如上所述，尽管对语言意义的获取有着种种难以逾越的困难，但是人工智能研究者们还是在语言的机器理解方面开展了大量的研究工作，这样的研究，统称为自然语言理解。

就汉语而言，目前的研究内容主要包括词语切分与标注、句法语篇分析、歧义消解等多个方面。为了突出意义理解这一核心问题，我们分别从意群分割、句法分析以及歧义消解三个方面来介绍自然语言处理主要问题及其处理方法。

1、多尺度意群分割

首先就是多尺度意群分割问题。

所谓意群，指的是我们的语言所表达的思想，都是通过一群相互关联一起的意义单元体现出来的，而这些意义单元根据其所处语言片段的角色，有大有小，

因此意群分割也就有一个多尺度问题。

1、多尺度意群分割

实际上，语言理解就是一个“依篇断句，析名分词”的过程。小到音节的切分，大到段落划分，无不贯彻着这样一个中心问题。

因此不管是用耳朵听读，还是用眼睛看读，这一过程的核心问题都是要根据语言的运用规律，层层分解不同尺度大小的语言单元，简称语元，然后在这些不同尺度层次的语元及其相互关系中来理解整个语篇的思想内容。

而这其中层层分解出不同尺度的语元，就是语言理解中的意群分割问题的任务。

1、多尺度意群分割

过去，我国文言文不用标点符号，句读成为断句破文的重要技巧。

例如，从前有一老汉老年得子，耽心死后女婿抢夺儿子的家产，于是留下遗嘱：“老汉八十生一子，人云非是我子也，家产事业均属予女婿，外人不得争执。”待儿子长大后，他按照老汉临终前的嘱咐，到官府告状说遗嘱原意是：“老汉八十生一子，人云非，是我子也；家产事业均属予，女婿外人不得争执。”结果老汉的儿子终于要回了应由自己继承的遗产。

1、多尺度意群分割

其实，不仅对于古代汉语意群分割中这种多重性解读十分突出，即使对于现代汉语，如果你着手的是语词分割问题时，那么由于再也没有标点符号和空格可资利用，这时同样也会存在多重性歧义分割问题。

特别是，原则上不同尺度意群分割之间存在着非常相似的规律，因此这种歧义性分割问题也是具有普遍性的现象。

1、多尺度意群分割

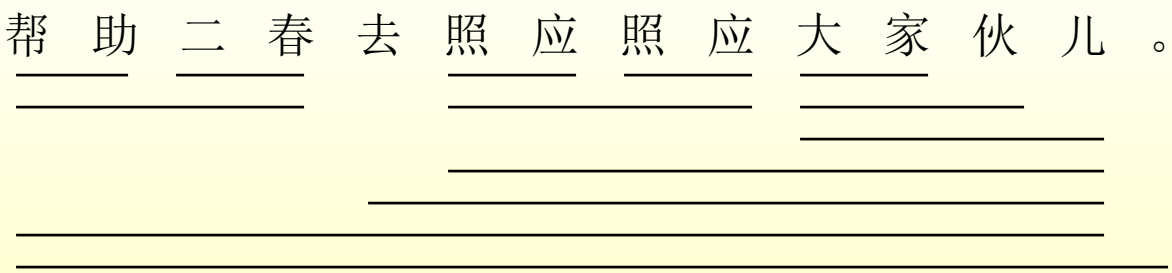
实际上，这样的认识在我国的语言学家朱德熙的《语法答问》中早就有过明确的论述。朱德熙认为汉语句子的构造原则跟词组的构造原则基本上是一致的，因此他明确提出了以语词为基点的语法体系。

对于意群分割问题，这就意谓着只要能够解决语词切分问题，那么就能够打下语法分析的基础，而其它尺度上的意群分割，也就可以通过语词意群的语法组合和语境制约来实现。

1、多尺度意群分割

那么，又如何进行语词意群分割呢？在语言理解研究中，语词意群分割的任务就是要从构成语句的语词层次开始，逐级地划分出全部语元并建立起反映层次关系的联系。

例如，对于语句“帮助二春去照应照应大家伙儿。”下图便是反映这种要求的意群分割结果。这其中最关键的便是一个个语词的切分问题。



1、多尺度意群分割

对于机器来说，如果我们已经建有一个完全的机器词典，其中收录了全部可用语词，那么最简单的机器语词切分方法就可以采用最大语词匹配策略来进行。

这种方法通过依次读取语句中的汉字，并当汉字串积累到最长且构成为机器词典中可接受的语词，那么就将该汉字串作为一个语词对待，然后再从语句余剩的汉字流中如法去分割下一个语词，如此等等，直到语句结尾，就完成全部语词级的切分工作。

对上面例子就是用最大语词匹配法来切分得到的结果。采用这种方法，对于比语词大的意群语元划分，则归于语法分析去完成。

1、多尺度意群分割

很明显，由于语词搭配多种可能选择的存在，这种最大匹配切分方法虽然能够保证切分出来的语词均是合法的，但却不能保证这种切分结果是在句法上是合理的。

例如对于“他的确切意图是什么？”按照最大匹配切分法结果为：

他丨的确丨切丨意图丨是丨什么丨？

而合理的应该是：

他丨的丨确切丨意图丨是丨什么丨？

尽管对于“他的确切菜了”能保证正确分为

“他丨的确丨切丨菜丨了”。

1、多尺度意群分割

语词切分的歧义性问题是普遍存在的。

例如：

“这个故事太｜平淡了”、

“请将｜军用毛毯盖在她身上”、

“美国｜会采取行动制裁伊拉克”、

“该研究｜所得到的奖金很多”等俯拾皆是。

而歧义性语词切分只有在一定的语境考察下才能够得到正确的切分，仅靠机械的最大匹配显然是不能彻底解决语词的切分问题的。

当然，为了利用语境上下文制约的关系，我们可以通过对所有可能切分作最优选择的策略来进行语词的切分。

1、多尺度意群分割

遗憾的是，有时词语关联的确定还会依赖于更高层次意义的理解，也就是说只有在理解了整个语句之后才能够确定语词的分割。

这样由于更高层意义的理解反过来无疑又是要依赖于分割好的语词的，于是就有一个语词分割与语句整体意义理解相互依存的问题。

因此，如果在这基础上，再进一步考虑跨层次相互作用问题，那么意群分割看似一个小问题，实际却是动一牵百的大问题，甚至与整个语篇的理解密不可分。

1、多尺度意群分割

于是，想要解决意群分割问题，我们就离不开意义的整合问题，而意义的整合问题，反过来又是以意群分割为基础的。

在语言理解的机器实现研究中，为了避免这种无谓循环，往往采用：

在一种初步的意群分割之后，
再考虑面向意义的句法分析。

End 1

2、依存性句法分析

如果说意群分割是一种自下而上的理解步骤，那么句法分析便是一种自上而下的理解步骤。

二种步骤的相互补充递进，也许就可以在某种程度上突破那种“无谓循环”的桎梏，走出语言理解的困境。

问题果真如此吗？那么机器又是如何面向语义来进行句法分析的呢？1968年，美国语言学家C.J.Fillmore提出的格语法无疑为此打开了一丝光亮。从而引发了一场面向语义进行语句分析的浪潮。

2、依存性句法分析

面向语义进行语句分析，就是要通过语句中语词关系分析，建立起语词之间的各种语义联系。

比如: 在“我开门”中，“我”是“开”的实施者，“门”是“开”被施者。在这个简单句子的理解中，你只有建立起“我”、“门”和“开”之间这种正确的语义关系，你才能够真正理解这一句子的意义。

而这种建立语义联系进而理解句子意义的能力是每一个语言掌握者所具备的基本能力。

2、依存性句法分析

从属关系语法的创始人特思尼耶尔(L.Tesniere)在《句法结构的要素》(Elements de syntaxe structurale)中就强调：“所谓造句，就是建立一堆词之间的各种关联，给这一堆词赋予生命；反之，所谓理解句子，就意味着要抓住把不同的词联系起来的各种关联。”

基于这样的原则，在格语法提出以后，作为一种直接发展的结果，1973年美国耶鲁大学的科学家C.Shank提出了一种概念依存语法理论，并在机器理解语言的研究中广泛被应用采纳。

2、依存性句法分析

概念依存语法的句法分析方法首先是找出语句中的主名词和主动词，形成一个初步的语义结构，然后通过在语句中寻找所需要的其它成份来不断完善这一语义结构，最后给出语句所反映出来的概念依存网络。

例如，对于语句“日本海沿岸多雪”，以“多”为中心词的概念依存分析可表示为：

日本 ←←← 海 ←←← 沿岸 ←←← 多 →→→ 雪

依存定语 依存定语 依存主语 依存宾语

这就给出了“日本海沿岸多雪”语句的语义结构。

2、依存性句法分析

看得出来，只要定义足够反映语义依存关系，那么通过确定概念依存关系，是能够有效地得出语句所对应的语义结构的，并表示为机器内部可以处理的形式。

概念依存理论注重的是语言成分之间的外部联系，强调语句中各成分之间存在的功能关系。

而这些功能关系又是以主名词或主动词为中心建立起来的，这样就突出了中心语词在语句语义结构中的中心作用。这便是中心词驱动语义分析理论的滥觞。

2、依存性句法分析

依存语法描述的是句子中词与词之间直接的句法关系。这种句法关系是有方向的，通常是一个词支配另一个词，或者说，一个词受另一个词支配.

所有的受支配成分都以某种依存关系从属于其支配者。这种支配与被支配的关系体现了词在句子中的关系。

2、依存性句法分析

按照依存关系来进行句法分析，一般遵循如下四个基本约定：

（1）一个句子中只有一个成分是独立的；

（2）其他成分直接依存于某一成分；

（3）任何一个成分都不能依存于两个或两个以上的成分；

（4）如果A成分直接依存于B成分，而C成分在句子中位于A和B之间，那么：C或者直接依存于A，或者直接依存于B，或者直接依存于A和B之间的某一成分。

2、依存性句法分析

根据上述约定，句子中有惟一一个独立成分，称之为中心语（可以是单个的词，也可以是由两个或两个以上的词组合成的短语）。它作为依存关系树的根节点，其他成分都依存于中心语。

这些成分有些对句子的结构起决定性作用，称为基本句型成分，包括主语、状语、补语、宾语（含第二宾语）。

还有些成分**独立于**句型结构，主要用于表示插话、句子的语气、时态或停顿等，称为附加成分，包括插入语、叹词、句末语气词、呼告语、应答语、动态助词和标点符号等。

2、依存性句法分析

在汉语的基本句型中，绝大多数句子的中心语是由动词（短语）担当的，只有少数句子其中心语是由形容词或体词担当的。

同样在汉语的基本句型中，绝大多数的句子的主语和宾语都是由名词（短语）担当的，只有少数句子其主语和宾语是由形容词或动词（短语）担当的。

由于句子的中心语支配着句子中的其他成分（主语、宾语、状语、补语），所以有必要对动词、名词和形容词的语义知识进行分析并加以分类，进而能从中总结出中心语与各被支配成分之间的语义关系。

2、依存性句法分析

这种表示方法简单、直观，比较接近人的分析句子方法，因此较适合人对句子的理解，同时比较便于转化为机器内部语义表示形式。

实际上，只要定义足够反映语义依存关系，那么通过概念依存关系的确定，是能够有效地得出语句所对应的语义结构的，并表示为机器内部可以处理的形式。

2、依存性句法分析

比如，在依存语法分析结果中，可以将**各对**关系用三元组表示，即(A，B，R)，其中A和B分别代表句子中的词语，R表示词语A与B之间的关系。R是个有向弧，它由B(支配者)指向A(被支配者)，即词语A的向上依存关系为R。

采用这种表示方法，具有以下特点：

（1）便于机器表示；

（2）与依存关系网的平面表示法一一对应；

（3）词语与词语之间的依存关系清晰、直观；

（4）充分体现了自然语言的不对称现象，较好地解决了自然语言语义结构表达式表达问题。

End 2

3、语境中意义获取

一旦通过依存句法分析获得了一个语句的全部依存关系，接下来就可以将其转化为某种逻辑语义结构的表示形式，以便机器内部表示并进一步开展其他的语义表达处理。

比如，适当地根据各种语义关系将一些词语转化为相应的谓词，那么就可以采用一阶谓词逻辑表达式来给出语句的意义表述。

3、语境中意义获取

当然上述句法分析的有效性给予这样一个前提，就是我们的语言是无歧义的。

我们十分清楚，对于严格的形式语言，机器可以很容易实现不同语言之间的相互转换。

因此，如果我们能够设计一种表示语义的形式描述语言，其满足无歧义性、具有简单的解释规则和推理规则以及具备由语句形式确定的逻辑结构，那么我们就可以通过定义良好的形式语言（比如某种逻辑系统）来确切地表述出给定自然语言语句的语义，只要自然语言的语句含义是无歧义的就行。

3、语境中意义获取

但事实上，没有哪种自然语言是不存在歧义现象的，甚至可以说歧义现象是自然语言的一种固有属性。

因此，要想解决语句的语义分析，从而可以用形式语言来描述自然语言语句的语义，就不可避免地要解决语言的歧义消解问题。

可见在机器的语言理解研究中，不仅不能忽视或回避语言的歧义现象，而且还应将歧义现象的研究和处理作为更重要的课题来研究。因此可以说，如果没有对语言歧义现象有全面深刻的认识和把握，要想实现机器的语言理解是不可能的。

3、语境中意义获取

通常语言歧义的表现形式有三类，即模糊、双关和选择。

模糊歧义是指语言表达内容时含糊不清而造成的歧义，往往是说写者连自己也不能明确所要表述的观念为何，或者是因无谓的同义反复、无益的语辞矛盾或明显的语焉不详所致。

例如“狗给狼吃了”（到底谁吃了谁？）、“明天晴天，别忘了带雨伞！”（晴天带雨伞，莫明其妙）、“王守信全家红”（是得彩了呢，还是成了红人？）之类，都是模糊歧义的例子。

双关歧义则是指多种意义的同时关联，融合成为或暗有所指、或一箭双雕、或蕴意深刻等效果的言辞表述。

例如推销皮鞋油的说：“一流产品，为足下争光”、情人的“藕断丝连”、歇后语“竹蓝子打水（暗指一场空）”等等。

3、语境中意义获取

选择歧义与双关和模糊歧义都不同，指的是语言中有多种明确独立相互排斥的意义选择可能，每种选择的意义并不模糊。

例如：“我们三个一组”（要么“我们丨三个一组”要么是“我们三个丨一组”）、

“女子理发店”（要么是“为女子理发的店”要么是“女子开的理发店”）、

“有的作品写年轻的妻子死了丈夫发誓不再结婚”（要么是“丈夫死了，妻子发誓”要么是“妻子死了，丈夫发誓”）、

“白头翁死了”（要么是人死了、要么是草死了、要么是鸟死了）等都如此，

有多种意义可供选择，并选择了一种就排除了其它选择可能。而每种选择可能的意义又都是明确无歧的。

3、语境中意义获取

由于能够左右歧义确认的外界条件主要是语境，因此对歧义语句的理解，只有联系语境才能正确把握其正确意义，而仅靠句子本身的结构成分及其组合意义是不够的。为了使机器也能够进行歧义消解工作，就必须有一种强调语境条件的语义分析方法。

美国语言学家J.Barwise和J.Perry提出的情境语义学正是基于这种要求而产生的一种语义分析新方法。跟传统语义的解释模式比较，新方法强调的正是语境条件的参与。这样对语句的理解就不仅仅是“意义表达”及其真假取值（指称真值）的结合，而是“意义表达”与在一定的语境条件（前提）参与下得出的真假取值的结合。

3、语境中意义获取

当然，大多数歧义是可以通过语言的或主观的语境条件来消除，这也是人类语言理解能力最有效、也最基本的机制之一。

因此，对于语言理解的机器实现而言，重要的不是寻找语境来使语言不含有歧义，而是要在给定的语境中来理解歧义的语言。自然语言的理解，说到底就是一种解释，将歧义的语句通过语境条件作用得出其尽可能确定的意义，或者同时保留多种关联或选择的意义，并用机器可以严格无歧义处理的形式加以表述。

3、语境中意义获取

用语境条件下的语义解释模式可以处理歧义语句的机器理解中的非模糊类的歧义问题，方法是对多种歧义理解可能分别用逻辑表达式表示，然后形成“与”（代表双关歧义）“或”（代表选择歧义）逻辑联式。最后，再利用语境条件式来推演，取这一逻辑联式中真假值为真的某个逻辑表达子式为其结果。

如果结果为真的子式不唯一，则表示在此语境条件下也不能完成歧义消解。

3、语境中意义获取

总之，语言意义的不确定性、对语境的敏感性，是自然语言最重要的功能表现，而歧义的产生也是诸种因素相互作用过程中处于不同动力学制约下状态冲突的、必然会出现的现象。

言传与意会之间、思想与现实之间以及语言表达能力与表达内容认识之间等等的不一致性，都会成为歧义产生的源泉。甚至我们还会有意地利用歧义现象来表达特定场合下的思想和情感，以达到一定的目的。

End 3