歧义消解策略初探

冯志伟

(国家语言文字工作委员会)

摘要:本文比较了中文和英文在歧义现象上的异同,分析了基于"制约"的歧义消解策略和基于"优选"的歧义消解策略,指出了自然语言的歧义结构本身就包含了消解歧义的因素,认真地分析这些因素,可以为歧义消解提供有用的条件,这些条件包括再分类、句法制约条件、语义制约条件等.

On strategy of disambiguation

Feng Zhiwei

(The State Language Commission)

Abstract: In this paper, the author compared the differences in Chinese and English ambiguity, proposed the strategy of disambiguation based on the constrant and the preference. The author emphasized that the ambiguous stracture includes the factors for disambiguation, these factors are : subcategorization, syntactic constraint, semantic constraint etc.

语言中的同形歧义既反映在单词上,又反映在由单词组成的各种结构上,形成词汇歧义(lexical ambiguity)和结构歧义(structural ambiguityn).

单词的兼类现象就是一种词汇歧义,兹不赘述。本文主要讨论结构歧义问题。

英语中的结构歧义,最常见的有如下三种:

(1)在"VP + NP1 + Prep +NP2"这样的结构中,介词词组 PP(Prep + NP2)既可以作为名词词组 NP1 的定语,又可以作为动词词组 VP 的状语,这就产生了歧义。

例如,句子"I saw a boy with a telescope"中的 NP2 "a telescope",当它作为 NP1"a boy"的定语时,句子的含义是"我看到了一个带着望远镜的男孩"(试比较:"I lost the ticket to Berlin"[我丢失了去柏林的车票]);当它作为 VP"saw"的状语时,句子的含义是"我用望远镜看见了一个男孩"(试比较:"I send the ticket to Berlin"[我往柏林寄出了车票])。

(2)当若干个词与 and 连用时,由于 and 的管辖范围不同,而影响到层次结构的不同。

例如,短语"old men and women"可解释为"年老的男人和所有的女人",这时,层次结构为((old men) and women),and 与 old 无关,也可解释为"所有年老的男人和所有年老的女人",这时,层次结构为(old(men and women)),and 与 old 有关。

(3)当两个或两个以上的名词组成词组时,对整个名词词组的含义往往可以作不同的解释,就会产生歧义。

例如,由名词 widget(作附件用的小机械)和名词 hammer(锤子)组成的名词词组 widget hammer,既可以理解为"widget used as hammer"(作锤子用的小机械),又可理解为"hammer

* 本项研究得到国家自然科学基金的资助,项目号为69483003.

for hitting widget"(锤击小机械的锤子),从而产生歧义;如果在前面再加上一个名词 town(城市),组成名词词组 town widget hammer,其层次结构可分析为((town widget) hammer),又可分析为(town(widget hammer)),这样的名词词组的歧义就更为严重了。

如果在一个英语句子中,既包含有"VP + NP1 + Prep + NP2"这样的结构,其中的 NP1 或 NP2 又是由若干个名词组合而成的名词词组,并且还包含连接词 and,那么,这个句子的歧义将成倍地增长,其剖析的难度也就更大了。

英语中第一种常见的结构歧义,即介词词组 PP 既作状语又作定语的那种歧义,汉语中并不多见。因为汉语的 PP 作定语时,一般置于名词词组之前,常加"的",不易与作状语的 PP 相混。但是,在汉语的介词词组中,由于介词管辖范围的不同,却容易引起歧义。例如,

关于((教师的)小说)

(关于(教师的))小说

在第一个短语中,介词"关于"的管辖范围是"教师的小说"(试比较:"关于动物的尾巴"),在第二个短语中,介词的管辖范围只是"教师"(试比较:"关于动物的书"),因而产生歧义。

英语中第二种常见的结构歧义,即由于连词 and 的管辖范围不同而产生的歧义,在汉语中也存在。在汉语中,"的"字跟连词"和"用在一起,最容易产生管辖范围的问题。例如,

把((重要的书籍)和(手稿))带走了

把(重要的(书籍和手稿))带走了

又如:

((车票)和(零用的钱))都在这里了

((车票和零用)的钱)都在这里了

英语中第三种常见的结构歧义,即由两个或两个以上的名词组成名词词组而产生的歧义,在汉语中也很普遍。

由名词 N1 和名词 N2 组合而成的词组,其结构关系各有不同,形成结构歧义。例如,

(N1) + (N2)

(女子)(理发店)

可以指专门给女子理发的理发店,也可以指理发师全都是女性的理发店。

由三个名词组合而成的词组,由于结构层次的不同,也会产生结构歧义。例如,

(N1 + (N2 + N3))	((N1 + N2) + N3)
(儿童(文学作品))	((儿童文学)作品)
(中国(历史研究会))	((中国历史)研究会)
(北京(大学毕业生))	((北京大学)毕业生)
(台湾(语言研究会))	((台湾语言)研究会)

由形容词 ADJ、名词 N1、名词 N2 组合而成的词组,结构层次不同,也会产生结构歧义。例如,

(ADJ + (N1 + N2))	((ADJ + N1) + N2)
(小(学生字典))	((小学生)字典)
(新(文学概论))	((新文学)概论)
(新(职工宿舍))	((新职工)宿舍)

事实上,汉语中常见的同形歧义结构还有许多,情况似乎比英语更为复杂。

在自然语言处理的研究中,早在 60 年代,美国哈佛大学教授久野(Kuno)就提出了歧义消解(disambiguity)的问题。

久野指出,英语句子

Time flies like an arrow

有若干个歧义的分析结果。因为 time 可以为名词(词义为"时间"),也可以为动词(词义为"测定、拨准"等),还可以为形容词(词义为"定期的"),flies 可以为动词现在时单数第三人称(词义为"飞"),也可以为名词复数(词义为"苍蝇");like 可以为动词(词义为"喜欢"),也可以为介词(词义为"如像")。这样,这些词可以组成结构各不相同的句子,形成歧义句。其含义分别为:

① 时间像箭一样飞驰;② 测量那些像箭一样的苍蝇;③ 定期飞来的那些苍蝇喜欢箭.

学者们普遍感觉到,歧义是语言自动分析的一个棘手问题。他们指出,如果使用如下的短语结构语法来分析英语,其歧义将十分严重。

S → NP VP	NP → Pron
VP → DET N PP	VP → V NP PP
PP → ε	PP → Prep NP

这个短语结构语法可以生成句子:

I saw a man with a telescope

这个句子中的 PP (with a telescope) 可以修饰名词 man,也可以修饰动词 saw,从而得到两种不同的结构,相应地得到两种不同的意义:"我用望远镜看到一个人","我看到一个带着望远镜的人"。

这个短语结构语法还可以生成句子:

I saw a man in the park with a telescope

这个句子中有两个 PP:一个 PP 是 in the park,一个 PP 是 with a telescope. 这两个 PP 可以同时修饰名词 man,也可以同时修饰动词 saw,也可以第一个 PP 修饰名词 man,第二个 PP 修饰动词 saw,也可以第一个 PP 修饰动词 saw,第二个 PP 修饰名词 man,从而得到四种不同的结构,相应地得到四种不同的意义:"我看到一个在公园中带着望远镜的人","我在公园中用望远镜看到一个人","我用望远镜看到一个在公园中的人","我在公园中看到一个带着望远镜的人"。从纯粹句法的观点来看,第一个 PP 中的名词 park,还可以被第二个 PP "with a telescope" 修饰,其含义为"装有望远镜的公园"(现在确实有些公园装了望远镜供游客观看远处风景之用),如果把这种情况也算进去,那么,还可以得到第五种结构,相应地得到第五种不同的意义。

有人指出,如果句子中的 PP 数目增加为三个,则可能得到的结构数目将会增加到十四个,相应地得到十四种不同的意义。

可以看出,在采用上述短语结构语法生成的句子中,当 PP 数目为 1 时,其歧义结构数为 2,当 PP 数目为 2 时,其歧义结构数为 5,当 PP 数目为 3 时,其歧义结构数为 14。如下所示:

PP 数	1	2	3	…
歧义结构数	2	5	14	…

可见,随着 PP 数目的增大,其歧义结构的数目以极快的速率增大,从而使自动句法分析的难度也迅速地增大,当 PP 数目很多时,如果要得出一切可能的歧义结构,自动分析的效率将会降低,并且可能发生组合爆炸的现象,导致系统的崩溃。

然而,从我们提出的"潜在歧义论"(简称"PA 论")可知,自然语言本身在具有潜在歧义的

词组类型结构(简称为"PT－ 结构")的实例化过程中,有自行消解歧义的功能,我们只要自觉地利用这种功能,就有可能达到部分地消解歧义的目的。

目前,在自然语言的计算机处理中,普遍采用的歧义消解方法,归纳起来不外两种:一种是基于"制约"(constraint)的歧义消解方法,一种是基于"优选"(preference)的歧义消解方法。

所谓基于"制约"的歧义消解方法,就是利用句法、语义制约条件,排除不能满足制约条件的结构,从而达到歧义消解的目的。

在 PT－ 结构实例化过程中,由于词汇单元之间句法条件的制约,往往能够消解歧义。例如,汉语中"数量结构 ＋ NP1 ＋ 的 ＋ NP2"这样的潜在歧义结构,可以解释为"(数量结构 ＋ NP1) ＋ 的 ＋ NP2",也可以理解为"数量结构 ＋ NP1 ＋ 的 ＋ NP2"。如果数量结构中的量词既能限定 NP1,又能限定 NP2,那就必定会产生歧义;但是,如果我们根据 NP1 及 NP2 的性质,对数量结构中的量词作进一步的再分类(subcategorization),使得数量结构中的这个量词不能同时限定 NP1 及 NP2,便可以消除歧义。

当这个 PT－ 结构实例化为"三个学校的实验员"时,由于量词"个"既可以限定 NP1"学校",又可以限定 NP2"实验员",因而不能消除歧义。

根据汉语的语法知识我们知道,"学校"的量词一般为"所","实验员"的量词一般为"位",据此我们对量词做再分类,把"学校"的量词规定为"所",将上述把 PT－ 结构实例化为"三所学校的实验员",由于量词"所"不能限定 NP2"实验员",其结构只能理解为"(三所学校)的实验员",歧义得到消解;我们如果把"实验员"的量词规定为"位",将上述 PT－ 结构实例化为"三位学校的实验员",由于量词"位"不能限定 NP1"学校",其结构只能理解为"三位(学校的实验员)",歧义也可得到消解。

采用这样的再分类的办法,不仅把量词分为若干小类,还可以把名词分为若干小类,把形容词分为若干小类,把动词分为若干小类,然后指出,哪些小类可以跟哪些小类组合,哪些小类不能跟哪些小类组合,便可以在潜在歧义结构实例化的过程中,利用这样的句法制约条件,达到消解歧义的目的。

除了再分类之外,还可以根据其他的句法关系来消解歧义。

在英语中,"Look at the pages of the book which are written by him"(看一看书中他所写的那几页)在结构上也有歧义,Which－ 从句"which are written by him"可能修饰 the book,也可能修饰 the pages。根据"从句中名词的数应该与被修饰的名词一致"这样的句法关系,从句中用 are written,是复数,故被其修饰的名词应该为复数,不可能是 the book,而应该是 the pages。根据这样的句法条件,歧义得以消解。

句法的制约条件有时显得过于烦琐,如果在 PT－ 结构实例化过程中利用词汇单元之间的语义制约条件,往往能够更加便捷地消除歧义。

例如,"VP ＋ 数量结构 ＋ NP"这个潜在歧义结构,其层次有时可以理解为"(VP ＋ 数量结构) ＋ NP",数量结构作 VP 的补语,有时可以理解为"VP ＋ (数量结构 ＋ NP)",数量结构作 NP 的定语。对于这样的潜在歧义结构,我们可以采用句法制约条件,对量词进一步作再分类,然后,说明哪些量词能与哪些动词结合形成述补结构,哪些量词与哪些名词结合形成定中结构,就可以进行歧义消解。但是,这样做比较烦琐,如果采用语义制约条件,根据语义上是否成立来判断能否形成歧义,从而达到歧义消解的目的,就显得更加便捷。例如,当实例化为"讲了三年历史"时,可以理解为"(讲了三年)历史","三年"作"讲了"的补语,表示讲历史讲了三年,也可以理解为"讲了(三年历史)","三年"作"历史"的定语,"三年历史"作"讲了"的宾语,表示讲了三年之内的历史,这时,潜在歧义转化为现实歧义。如果把"三年"换成"三千年",实例化为"讲了三千年的历史",则只能理解为"讲了(三千年的历史)","三千年"

只能理解为“历史”的定语，而不能理解为“讲了”的补语，因为从语义上来看，“讲了三千年”在语义上是荒谬的。这样，只需把“三年”换成“三千年”，便可以直截了当地消解歧义。由此可见，使用语义制约条件的便捷之处。

自然语言处理中普遍采用的另一种歧义消解的方法是基于“优选”的歧义消解方法。

所谓“优选”，就是在若干个存在歧义的候补结构中，选出一个最优的结构，从而达到歧义消解的目的。

对于具有潜在歧义的若干个候补结构，可以根据候补结构的优先度来进行优选，消解歧义。

例如，汉语中的“N + V + NP + AP”这个潜在歧义结构，其层次可以解释为“(N +V + NP) + AP”，是一个以小句为主语的主谓结构，又可以解释为“(N) + (V) + (NP + AP)”，其中的“(V) + (NP + AP)”是一个述宾结构，又可以解释为“(N) + (V) + (NP) + (AP)”，其中的“(V) + (NP) + (AP)”是一个兼语结构，这样，“N + V + NP + AP”便具有主谓(以小句为主语) －－ 述宾 －－ 兼语潜在歧义。

海外学者 Chao－Huang Chang 和 Gilbert K. Krulee 根据中国人讲汉语时的语感指出，在这样的潜在歧义结构中，逻辑主项(argument reading) 的结构应该优先于逻辑附加项(adjunct reading) 的结构。兼语结构和述宾结构都是属于逻辑主项的结构，而以小句为主语的主谓结构，其谓语为 AP，AP 是逻辑附加项，因而应该属于逻辑附加项的结构。这样，兼语结构和述宾结构的优先度应大于以小句为主语的主谓结构的优先度。当出现歧义时，应该优选兼语结构和述宾结构，从而达到消解歧义的目的。

这样，当 PT－ 结构“N + V + NP + AP”实例化为“张三笑李四很笨”时，可以理解为“张三 / 笑李四很笨”，“笑 / 李四 / 很笨”是一个兼语结构，又可以理解为“张三笑李四 / 很笨”，这是以小句“张三笑李四”为主语的主谓结构。根据兼语结构的优先度应大于以小句为主语的主谓结构的优先度的原则，应该选取兼语结构，排除以小句为主语的主谓结构。

当实例化为“小王说故事很有趣”时，可以理解为“小王说 / 故事很有趣”，“说 / 故事很有趣”是一个述宾结构，也可以理解为“小王说故事 / 很有趣”，是一个以小句为主语的主谓结构。根据述宾结构的优先度应大于以小句为主语的主谓结构的优先度的原则，应该选取述宾结构，排除以小句为主语的主谓结构。

根据说话人的语感来规定结构的优先度并不是很科学的。在上面的例子中，把“张三笑李四很笨”中的“笑 / 李四 / 很笨”理解为兼语结构，把“小王说故事很有趣”中的“说 / 故事很有趣”理解为述宾结构，并不能绝对地排除把“张三笑李四很笨”和“小王说故事很有趣”理解为以小句为主语的主谓结构的可能性。因为语感上的优先度只是表明了某种选择的可能性，并不能绝对地表明这种选择的合理性和现实性。语感上的优先度往往有着强烈的主观色彩，常常因人而异，难免有见仁见智之弊。

为了保证优选的客观性，莱斯克(M. Lesk) 提出利用既存的知识源来进行优选。机器可读词典中词典条目的定义是一种既存的知识源，当判断两个单词之间的亲和程度时，可以比较这两个单词在机器可读词典的定义中同时出现的词语的情况，如果在两个单词的定义中都出现共同的词语，便可推断它们之间的亲和程度较大，从而据此来进行优选。

例如，在英语中，pen 是一个多义词，可以理解为“笔”，也可以理解为“动物的围栏”，如果在一个句子中既有 pen，又有 sheep，而在机器可读词典的 pen 的定义中有“an enclosure in which domestic animals are kept”，在 sheep 的定义中有“There are many breeds of domestic sheep”，在这两个定义中都存在共同出现的单词 domestic，从而可以判断，在这个句子中，pen 的含义应该是“动物的围栏”，而不是“笔”，从而消解了歧义。

詹森(K. Jensen)和比诺特(J－L. Binot)利用联机词典中的单词的定义来消解英语介词的功能歧义。

例如，英语的 with 这个介词，其功能可以表示 INSTRUMENT(工具)，又可以表示 PART－OF(部分 － 全体)关系，这就出现了功能上的歧义(case ambiguity)．在英语句子"I ate a fish with a fork"中，fork(叉子)的定义为"an instrument for eating food"，其中的 instrument 与 with 的功能 INSTRUMENT(工具)相同，故可判断 with 在这个句子中的功能应该是 INSTRUMENT(工具)，故此句的含义应该为"我用叉子吃鱼"。

在英语句子"I ate a fish with bones"中，bone 在机器可读词典中的定义是"a part of animal"，在 fish 的定义中，有"a kind of animal"，这与 with 的功能 PART－OF(部分 － 全体)关系相同，故可判断 with 在这个句子中的功能是 PART－OF(部分 － 全体)关系，这样，这个句子的含义应该是"我吃带骨的鱼"。

在实际的自然语言处理系统中，常常把基于"制约"的歧义消解方法和基于"优选"的歧义消解方法结合起来，用基于"制约"的方法排除那些不能满足制约条件的歧义，用基于"优选"的方法比较各种歧义的优先度，选取其中的最优者，从而达到歧义消解的目的。

自从 80 年代马丁·凯依(Martin Kay)提出功能合一语法(Functional Unification Grammar) 以来，在自然语言处理系统中普遍采用复杂特征集和合一运算的方法。人们发现，在自然语言分析系统中，随着分析的进行，包含自然语言中的信息是单调递增的，这就是自然语言分析系统中信息的单调递增性(information monotonicity)。

根据这种信息的单调递增性，有的学者提出，对自然语言分析过程中出现的歧义，应该做渐进的评价(incremental evaluation)。 有的学者提出了"渐进歧义消解法"(incremental disambiguation)。他们主张，当出现歧义时，不要匆忙地作出评价，等到自然语言分析系统中的信息单调递增到可以对这种歧义进行判断时，再作出判断，从而消解歧义。

在 PT－ 结构实例化过程中，由于词汇单元的插入，其信息也是单调递增的，因此，PT－ 结构实例化过程也具有信息的单调递增性，我们同样可以采用渐进歧义消解法。在信息不充分条件不成熟时，不必匆忙地消解歧义，等到信息单调递增到足以满足各种制约条件和优选的标准时，才进行歧义的消解。

在自然语言处理中，同形歧义的自动消解还是一个未彻底解决的问题，还有待我们做更深入的探索。

参考文献

[1] 冯志伟，中文科技术语的结构描述和潜在歧义，《中文信息学报》，1989 年，第 2 期。

[2] 孙茂松、黄昌宁，汉语中的兼类词、同形词类组及其处理策略，《中文信息学报》，1989 年，第 4 期。

[3] Chao－Huang Chang, Gilbert K. Krulee, Resolution of Ambiguity in Chinese and Its Application to Machine Translation, Machine Translation, 6, 1991/1992.

[4] 石安石，语义论，商务印书馆，1993 年。

[5] M. Lesk, Automatic Sense Disambiguation Using Machine Readable Dictionaries, In Proceedings of ACM SIGDOC Conference, 1986.

[6] K. Jensen, J－L Binot, Disambiguatng Prepositional Phrase Attachment by Using On－Line Dictionary Definitions, Computational

[7] 朱德熙，汉语句法里的歧义现象，《中国语文》，1980 年，第 2 期。