汉语自动分词中的歧义问题

侯 敏　　孙建军　陈肇雄

（黑龙江大学）　（中国科学院计算所机译中心）

摘　要： 汉语自动分词中的歧义现象，可以从两个不同的角度分类。从结构形式看，歧义字段可以分为交集型、组合型、混合型三类。从消除歧义的条件看，歧义字段可以分成后字有定型、语段歧义型、句子歧义型三类。而后一种分类对解决实际问题更有用。后字有定型可以用后字确定法及有条件后字确定法处理，语段歧义型可用上下文确定法处理，句子歧义型则用词频统计、主题确定等方法处理。它们要在不同的平面上实现。

Ambiguities in Automatic Chinese Word-Segmentation

Hou Min

(Heilongjiang University)

Sun Jianjun and Chen Zhaoxiong

(Institute of Computing Technology, Chinese Academy of Science, Beijing)

Abstract

The ambiguous phenomena in automatic Chinese word-segmentation can be classifiedfrom two different angles. As viewed from the form of structure, ambiguous character strings can be classified into three kinds: the overlapping type, the combinationaltype and the mixed type. And from the point of view of the conditions used to eliminate ambiguities, they can be classified into such three kinds as the back-character-being-definite type, the string-being-ambiguous type and the sentence-being-ambiguous type. Perhaps the latter classification is more conducive to solve the problems. The back-character-being-definite type can be processed by the method of specifying the back character(s), sometimes additional conditions need to be given. The string-being-ambiguous type can be handled by the method of giving the context within the sentence. The sentence-being-ambiguous type may be treated with the method of statistics of word frequency or determination of topics, etc. They are to becomplementedatdifferent levels.

引言

随着中文信息处理研究的不断深入，出现了多种多样的自动分词方法。评定一种切分方法或一个分词系统好坏的标准不外两条：一是速度，二是精度。而第二条尤为重要。要想提高切分精度，除了要建立一部（或几部）较完备的词典外，最重要的就是如何处理好切分中的歧义问题了。

为不同目的而做的分词系统很难用一种方法去实现，当然也不能用一个标准去衡量。一般的分词系统都是汉语自然语言处理系统中的一个子系统，分词系统和其它系统共用一部多功能汉语电子词典，分词和查词典同步完成。因此分词的目的是为了进行下一步的处理，不象有些为了统计词频而做的分词系统那样，要严格划清词与非词的界限。例如我们把"有时候"分为一个词，完全是从实用角度的便利出发，可以不考虑从理论上说它是不是词。一般分词采用的是正向直接匹配法，即当正向扫描到某字时，用词典中以该字为首字的词条去匹配素材。词典中只要求有相同字的词条要按先长后短的顺序排列。如遇歧义切分问题，则利用词典中的语法语义信息及歧义处理规则予以解决。

歧义问题的分析和处理

人们一般都把汉语自动分词中的歧义现象分为交集型、组合型两种。但就我们的实践来看，问题并不这么简单。就结构形式而言，统而分之，也至少可以分出交集型、组合型、混合型三种。但它们内部构成歧义的方式不同，语言特点不同，处理的方法也不尽相同。下面分别论述。

1. 交集型

交集型歧义字段是指分词中碰到的这样的现象：字段ABC 中，$AB\in W\wedge BC\in W$ (A、B、C分代表字串，它们多数情况下是一个字，也可以是两个或多个字；W是词的集合)。其中B是交集字串，交集字串的个数称为"链长"。如字段ABCD中，$AB\in W\wedge BC\in W\wedge CD\in W$，其中B、C都是交集字串，我们则说这个交集字段中交集字串的链长为2 。

在处理歧义切分问题时，首先应该明确：交集现象与歧义切分并不等同，这是两个概念。也就是说，交集字段不一定都会产生歧义切分或错分现象。至于哪些交集字段会产生错分，这与你使用的切分方法有关，或进一步明确地说，与你切分方法的扫描方向有关。如交集字段"对半导体（的研究）"，如果正向扫描，则可能产生错分，分成"对半/导体"；逆向扫描，则不会分错，切分结果自然是"对/半导体"。因为一般的系统中分词用的是正向匹配的方法，所以象下面(1)中这样的交集现象不会产生错分，就可以不去考虑了。

(1)1．实现在情报工作方面的自动化。

2．昨天下午他来了。

其中"实现在"和"昨天下午"分别是链长为1和2的交集字段。它们之所以是非错分型的，主要是因为ABC(D)组合在一起时，词A和词BC在一个句子中组合的概率几乎是零，如"实/现在"、"昨/天下"等在语言中是不成立的。所以，正确的切分是唯一的，只能是AB/C(D)。用正向直接匹配法，自然地切分为"实现/在"、"昨天/下午"。我们还发现，链长为2的交集字段有很多都属非错分型交集字段。理由同上。如(2)。

(2)1．我们可以正确切分这个句子。　　3．白天气温很高。

2．现在我们来量一下身长。　　4．确实现在物价很平稳。

使用正向直接匹配法可能出现错分的交集字段有以下几种情况。

1.1 后字有定型交集字段

(3) 1a．他只会诊断一般的疾病。　　2a．按时下的风气，一百元的礼金不算多。

1b．医生们在会诊时提出了新的方案。　2b．工人们按时下班。

其中的 "会诊断"、"按时下"等均是交集字段。正向切分，1a、2a都会错分成AB/C。仔细分析就会发现，在这类交集字段中，决定如何切分的是字段内的后字 C。也就是说，切分到AB时，还不能断定这样切分是否正确，要根据后字C才能确定。而就语言事实来看，切成 A/BC时的C是有定的。如1a中的 "会诊断"应切分成 "会/诊断"，"会"是能愿动词，它要求后接动词，而能出现在 "诊"后和 "诊"构成动词的字是有定的，只有 "断、疗、脉、治"等几个。 据此，我们在词典中 "会诊"词条下给出这样一条歧义处理规则：

会诊/断|疗|脉|治 —> @会@[诊 + N2]

其中 "/" 为项的分隔符，"|" 表示 "或"的关系，"@" 表示调用，"N2" 代表第2 项。这条规则的意思是：如果当前词 "会诊"后是 "断"或 "疗"、"脉"、"治"等字时 ，则要重新切分，将 "会"单独切为一个词，"诊"与第2项的字合成一个词，同时分别调取这两个词的所有信息 （因为分词和查词典同步完成）。否则就按原切分不变。这样1a、1b 都得到了正确切分。

这种用着字段内后字来确定切分的方法我们把它叫做 "段内后字确定法"，简称为"后字确定法"。

同样，用这种方法，我们也可以正确切分2a，但在切分2b时就出现了问题。 因为2b的正确切分不是 "按/时下/班"，而是"按时/下班"。其实，这是一个链长为2的交集字段，与上面例句 (2) 不同的是，其中的A(按) 和BC(时下) 在句子中可能组合，如2a。由此看来，后字确定法仅仅看到后字C、将BC组合在一起还是不够的，还必须延伸到D，即还应到 "BC" 这一词条下，看它是否还有歧义切分规则，如有，还应进行 "CD" 的匹配，匹配上了，应还原成AB/CD的切分。否则，保持A/BC不变。加上这一道手续， 后字确定法才是比较完整的了。通过这样的处理，"按时下的风气"和 "按时下班" 都可以正确切分了。

后字有定型交集字段的特点是：能与交集字B组合成词的后字 C是可以枚举的；前字A能单独成词；A与A前的字不能组成词；可用后字确定法在词这一平面处理。

1.2 语段歧义型交集字段①

这类交集字段的特点是：它的正确切分不取决于段内后字C，而取决于这一语段外的上下文，即句子中的其他因素。也就是说，就这一语段来说，切成AB/C或 A/BC都成立，它本身是歧义的。请看 (4)。为了分析的方便，我们不妨把可能出现的句子多罗列一些。

(4) 1a．他的确切地址在这儿。　　2a．他的确切菜了。

1b．谁能知道他的确切的意图是什么呢？　2b．菜的确切得不错。

1c．他语言表达的确切、生动是有名的。　2c．这次他的确切破了手。

1d．谁要是能知道他的确切想法就好了。　2d．他当时切没切肉？

— 他的确切了。

(4) 中交集字段 "的确切"本身既可以分成 "的/确切"，也可以分成 "的确/切"。但在具体的句子里，它只能取一种切分。在1组句子中只能切成"的/确切"，在2组句子中要切成 "的确/切"。它的歧义只停留在语段ABC这一层面，靠句内上下文可以确定它的正确切分。分析

上面的句子，可以看出，要做A/BC式切分，即 "确切" 是一个词，那么作为形容词，它后面只能接表抽象意义的CX类名词、判断动词VI、顿号或连词C。因为我们在词典里已给出了足够的语法语义信息，所以也可以正确切分这类句子。但在做法上由于它要看后面的语法语义信息，而这时后面的词还没切分出来，所以不能象后字确定法那样在第一次扫描时就处理好，而要在第二次扫描即处理短语时解决。也就是说，在第一次扫描时，先切成 "的确/切"，然后在 "的确" 词条下的短语规则中做这样一条规则加以校正：

的确/切/(的)/CX|VI|、|C —> @的 @确切 N4

其中 "( )" 表示括号内的各项为可有可无的任选项。这条规则的含义是，当 "的确" 和 "切" 两词相连且后面（有 "的" 可越过）又有抽象名词或判断动词或顿号或连词时，则重新切分为 "的" 和 "确切"，并调取这两个词条的全部内容，第3项删除，作为条件的第4项照样留下。这样，1组中的句子都可正确切分成 "的/确切"，而2组的句子不符合这条规则，仍保留 "的确/切" 的分法。

这种靠句中上下文来确定切分的方法我们叫它 "句内上下文确定法"，简称 "上下文确定法"。这种方法是在短语层面上解决歧义切分问题。

1.3 句子歧义型交集字段

这类交集字段在语言事实中很少出现，却是目前最难处理的歧义现象之一。它不仅交集字段ABC的切分是两可的，而且由此而及的整个句子也是歧义的。只有把句子放到更大的语境中才有可能作出它的正确切分，消除歧义。如(5)。

(5) 1. 请将军用毛毯盖在她身上。

2. 白天鹅在水里游来游去。

(5)1既可以切分为 "请/将军/用/毛毯/盖/在/她/身/上"，也可以切分为 "请/将/军用/毛/盖/在/她/身/上"。同样，(5)2中的 "白天鹅" 也有 "白/天鹅" 和 "白天/鹅" 两种分法。只有在具体的语境中，它们的意义才可能是确定的。对这种歧义切分现象，只有超越句子，用词频统计或主题确定等方法才有可能解决。

2. 组合型

组合型歧义字段是指字段AB中，$A \in W \wedge B \in W \wedge AB \in W$。凡组合型的歧义字段都不可能在第一次扫描中确定切分，要留待短语层或句子层处理。根据消歧条件的不同，组合型歧义字段还可分为两种。

2.1 语段歧义型组合字段

这类歧义字段占组合型的大部分。它的特点是歧义现象只停留在语段AB这一层面上。在具体的句子中，它或分或合，只有一种切分，即句子是单义的。如(6)。

(6)1a. 他将来上海工作。　　2a. 两个人一起去。

1b. 将来上海一定会更繁荣。　　2b. 这只不过是个人问题。

3a. 我从马上下来。　　4a. 现在差十分九点。

3b. 他马上就来。　　4b. 他十分高兴。

这些歧义字段，都可在句中找到条件确定它们的切分，即可以用句内上下文确定法处理。

1a中的 "将来" 应切开。"来" 作为一个表示带有趋向性动作的动词，它前面要求有一个表人的RL类或表组织的ZZ类名词作主语，带上副词 "将" 后，后面一般要有一个表地点的DD类名词作宾语。据此，我们在 "将来" 一词的短语规则中做一条规则：

RL|ZZ/将来/DD —> N1 @将 @来 N3

1a符合这条规则，"将"和"来"分开了；1b不符合这条规则，"将来"作为一个词保留来。下同。

2a中的"个人"做规则如下：

S/个人 —> N1 @个 @人

中S是数词，即数词后的"个人"须切开。

3a中的"马上"做规则如下：

P/(D S L J)/马上 —> N1 N2 @马 @上

中P表示介词，D是指示代词，L是量词，J是形容词。即介词后、中间可越过D、S、L、J的马上"须切开。

4a中的"十分"做规则如下：

十分/!J&!XV —> @十 @分 N2

中"!"表示"非"，"XV"表示心理活动动词。如果"十分"后面不是形容词或心理活动词，须切开。

这样，用上下文确定法，a组和b组的句子都得到了正确的切分。

2 句子歧义型组合字段

这类歧义字段与句子歧义型交集字段只是字段构成形式不同，性质却完全一样，也是目较难处理的歧义现象。如(7)。

(7)1．今天学生会讨论这个问题。

2．该研究所得到的奖金很多。

中的"学生会"可以分成"学生/会"，也可以不分，表示一个组织"学生会"。2句的"研所"也是如此，而且由此而及的句子意义也是两可的。这类句子象句子歧义型交集字段一，也只能用词频统计或主题确定等方法，在句群或篇章层面上消除歧义，确定切分。有的至在句群或篇章中也无法确定它的确切含义。

混合型

这种类型的歧义字段在语言事实中并不少，它集交集型与组合型的特点于一身，而且情形复杂。就我们目前接触到的语言事实来看，都是交集型内含组合型的，即交集字段的字长于组合字段。如(8)。

(8)1a．这篇文章写得太平淡了。

1b．这墙抹得太平了！

1c．即使太平时期也不应放松警惕。

2a．我们学会了解答问题的办法。

2b．他还不了解答题的方法。

2c．他学会了解方程。

2d．我们都了解他。

3a．乒乓球拍卖完了。

3b．你的乒乓球拍坏了。

析1组句子可以看出，1a中的"太平淡"是交集字段，1b、1c中的"太平"又是组合字段。交集字段长于组合字段，所以是交集字段中包含组合字段。那么，在处理这类歧义字段的候，也必须分两步走。第一次扫描时首先处理交集字段，如匹配成功，切成A/BC，如1a，问题解决；匹配不成功，如1b、1c，把AB保留下来，在短语层面再按组合字段处理。下面步说明。

3.1 第一步 ---- 交集字段的处理

有些混合型歧义字段这一步的处理可以用段内后字确定法。如1a中的"太平淡"。因为"太"是程度副词，它只能修饰形容词。而能和"平"构成形容词的只有"淡"、"常"、"凡"静"等字，据此，我们可以写出这样一条规则：

太平/淡|常|凡|静 —> @太@[平 + N2]

1a符合规则，得到了正确切分；而1b、1c不符合规则，"太平"保留下来，在第二步再按组合字段处理。

由于混合型歧义字段中的A和B各可以是词，所以较之单纯的交集字段有更大的灵活性，情况更复杂。有时即使是同一个后字C，切分也可以不一样。如2a、2b中的歧义字段"了解答"在2a中应切成A/BC，在2b中应切成AB/C。所以单纯用后字确定的算法就解决不了这类交集字段的问题了。但究竟切成AB/C，还是A/BC，还是有规律可循的。我们不妨把可能出现的句子多罗列一些。见(9)。

(9)1a．我们已学会了解答问题。　　2a．他了解答题的方法。

1b．乡亲们送走了解放军。　　2b．我们还不了解救人的英雄。

1c．人们高高兴兴地迎来了解放。　　2c．我们还不了解气体的性质。

1d．他知道了解毒的方法。　　2d．正是我对他的了解决定了我这样做。

1e．这位姑娘给我们作了解说。　　2e．他还真的不了解手的用处。

1f．他终于得到了解脱。　　2f．欧洲人还不了解围棋。

分析(9)会发现，1组中的歧义字段都应切成A/BC。其中"了"是动态助词，它要求前面必须是个动词。而2组则不然，AB应切在一起，"了解"作为一个动词，它前面一般不能出现动词。这样我们就可以在"了解"词条下做出这样的规则：

V/了解/答|放|放军|毒|说|脱|决|围|救|气|手 —> N1 @了 @[解 + N3]

其含义是，如果该词前有V，后为"答"或"放"等字，那么"了"字单切，"解"与后字合为一词。(9)1符合规则，"了"字单切；(9)2不符合规则，"了解"保留了下来。同样，(8)2a符合规则，得到正确切分；(8)2b、2c、2d不符合规则，留待第二步解决。这种处理方法比单纯的后字确定法多了一道手续，即前溯，往前看一个词，实际上就等于比单纯的后字确定法多了一个条件。所以我们叫它"有条件后字确定法"。因为我们用的是正向直接匹配法，前面的词已切分完毕，所以这种向前看的有条件后字确定法的实现不存在问题。

这样，(8)中1a、2a的问题解决了。

3.2 第二步 ---- 组合字段的处理

混合型歧义字段如果在第一次扫描时没切成A/BC，而是AB切在一起，那么切分并没有结束，因为AB还可能是A/B两个词。如(8)中的1b、2c。所以在短语层面上我们还要继续处理。

分析(8)1b，"太/平"作为形容词性偏正短语，出现的位置是有限的，只能作谓语或补语。作谓语时它能陈述除了"国家"、"单位"等组织类以外的名词，作补语必须用在"得"的后面。因此，我们在"太平"词下作一条规则：

N&!ZZ|得/太平 —> N1 @太 @平

这样用上下文确定法处理，(8)中的1b、1c都能正确切分了。

分析(8)2c，"了"是动态助词，它必须前附动词，又因为"解"也是动词，所以"了"前附的动词是有定的，必须是能带动词性宾语的谓宾动词V4。而"解"能带的宾语也是有限

的，只有算术题、方程等课程类名词KC，以及扣子、绳子等表具体事物的名词JT。所以我们可以在"了解"词下作一规则：

V4/了解/KC|JT —> N1 @了 @解 N3

这样(8)中的2c符合规则，"了"、"解"切开；而2b、2d不符合规则，"了解"是一个词。

(8)3a、3b属句子歧义型混合字段。对它们也只能超越句子用词频统计或主题确定等方法来处理，在句子内部无法解决问题。好在一、这类句子在语言事实中极少，即使不处理也不影响一般的实用化使用；二、就单个句子而言，做两种切分中的任何一种切分都可以说是正确的，所以一般不会出现错分问题。

结　　论

综上所述，汉语自动分词中的歧义切分现象可以从两个不同的角度分类。首先，从结构形式看，歧义字段可以分为交集型、组合型、混合型三类。它们的出现频率是交集型最多，组合型、混合型也不少。其次，从消除歧义的条件看，歧义字段则可以分成后字有定型、语段歧义型、句子歧义型三类。它们的出现频率是：后字有定型、语段歧义型都不少，句子歧义型则极少。这两种分类各有所长。不过，从解决问题的角度出发，后一种分类更有用。后字有定型的歧义字段可以用后字确定法及有条件后字确定法处理，语段歧义型则可以用上下文确定法处理，句子歧义型只能用词频统计或主题确定等方法处理。其中后字确定法及有条件后字确定法是在词这一层面实现，上下文确定法是在短语这一层面实现，而词频统计法等要在句子甚至篇章的平面上实现。

附注：①

"语段"是一个界限较模糊的概念。本文之所以用它，实在是不得已。因为象"的/确切"、"了/解"之类，很难说它是一个短语，而"了解"又是一个词。所以无法用一个界限明确的词语表述它。好在我们可以在这里限定一下：文中的"语段"是指歧义字段"ABC"或"AB"这一语言片断。它的身分可能是词，也可能是短语，还可能是一个无结构关系语段。

参 考 文 献

[1].梁南元等 <<汉语自动分词综述>> 计算机应用与软件 1987.9

<<汉语计算机自动分词知识>> 中文信息学报 1990.2

[2].张　普 <<现代汉语"有穷多层列举"自动分词方法的讨论>> 计算机与语言(3)

[3].何克抗等 <<书面汉语自动分词专家系统的设计原理>> 中文信息学报 1991.2

<<书面汉语自动分词专家系统的实现>> 中文信息学报 1991.3

[4].王开铸等 <<基于短语结构文法的分词研究>> 中文信息学报 1991.3