第18卷 第2期 自 动 化 学 报 Vol. 18, No. 2
1992 年 3 月 ACTA AUTOMATICA SINICA Mar., 1992

二维扩展属性文法中的三种识别控制机制

赵 明

(中国科学院软件研究所,北京 100080)

摘 要

本文介绍了识别手写印刷体汉字的二维扩展属性文法 (2-D EAG) 方法中多义文法、共生文法和结构推断三种识别控制机制。 采用这些控制机制，2-D EAG 方法可以在较大的幅度内容纳结构畸变,可以利用相似结构之间的类比,实现先外后内的识别顺序，以及抑制冗余识别.

关键词: 汉字识别,二维扩展属性文法,汉字结构类比.

一、引 言

在汉字识别的结构方法中,特征结构的存在与否是至关重要的.尽管可以采用一些模糊度量的方法,但一般是在划分了特征结构之后,因而只对克服基元级的畸变有效。 而特征结构的划分本身就是很不稳定的。 相比之下,对于人类识别来说,根据哪些特征结构辨认一个汉字,是在具体的识别过程中完成的.汉字中最大的信息冗余存在于部件的组合关系中.设有 100 个汉字部件,那么仅按上下、左右关系组合一次就可构成 20000 个汉字.实际上,大多数组合关系是不存在的,这就给人类书写带来了相当大的自由度。 例如,“氵”与“讠”易混,因此写“汗”时就得注意一些,以免与“讦”混淆。 但写“汉”时则比较随便，因为不存在混淆的危险。 这种利用相似结构之间的类比进行识别的方式对人类识别来说是极为自然的。 但对于机器识别,由于只是将输入与样板一一比较,因而难以利用这种样板之间的相似结构类比。 即使硬要度量样板之间的类似程度，也只能使用经过若干次变换后得到的一些数值(统计方法)或编码(结构方法)。 但这种度量已经丧失了原有的结构上的意义了.

对统计方法来说,结构越复杂的字有时越容易识别,因为较多的结构把输入与其他样板之间的距离拉大了。 对于结构方法,字的结构越复杂,需要做的判断就越多，判断失误的可能性就越大。 但对汉字结构的研究表明，结构复杂的字中有很多信息是冗余的。 这些冗余信息对区分一个汉字并没有多大用处,反而会带来不利影响。 例如“藏”字,即使没有内部的臣,也不会使人发生误识。 机器识别若硬要去辨认内部的“臣”,反倒有可能导致识别失败。 但到底有多少特征结构就能识别一个汉字？ 如果事先指定，也会引出新的问

本文于 1988 年 6 月 18 日收到.

题.例如,若识别出“王”、“田”,则足以识作“理”.但若右上角辨认不清,一时还难以区分是“田”还是“日”，那么这个字还有可能是“琨”。所以有多少特征结构就能识出一个字实际上是动态的，在不断排除不可能候选的情况下，当候选唯一并且满足某些认定条件时才可以说识出了这个字。为了利用汉字字形中的这一特点,减少不必要的判断和归约，2-D EAG 方法[1]中提出了多义文法、共生文法、结构推断三种识别控制方式。这三种方式利用了扩展属性文法[2]中的继承属性，具有很强的上下文处理能力。

二、多义文法

汉字集合中存在着大量的相似结构,如{日,曰}、{人、八、入}、{己,已,巳}等等。这些相似结构在某些情况下需要加以区分,在大多数情况下则不必区分。对于手写汉字来说,当不需要区分时往往很难辨认到底是哪一个。这种书写上的灵活性给结构识别方法带来很大的麻烦.它迫使识别机硬去做那些本来没有必要去做的区分,或者靠多样板来容纳各种变形。为了反映和利用汉字集合中的这一特点，2-D EAG 中提出了多义文法的概念及其实现算法。所谓多义,就是一个文法结点代表几个相似的部件,归约时当作一个部件处理。当不需要对它们作出区分时，它们看起来象一个部件一样。当需要作进一步区分时,则启动某个多义基础部件的多义区分函数,对多义进行区分。

例：汉字“汆,汆,余”中均出现𠆢形。“汆”中的必须是“人”,“汆”中的必须是“入”,而“余”则无所谓(实际上写成余,余,余的都有)。2-D EAG 中这些汉字的定义为

CODE, comb, SON1, SON2, dist, 〈PG list〉;

人,　　111, 〈八,入〉;

汆, 1*UD*, 人, 水, 101, 〈0, 汆〉;

余, *UD*, 人, 　0;

有〈　〉的 EAG 定义是一个多义文法,其中 CODE 项作为多义文法的代表名,同时也作为第一个多义选择项,〈　〉内是其他多义选择项，comb 项规定了组合方式（LR 表示左右结构，UD 表示上下结构)，dist 项是区分标记字位图，1 表示要区分，因此 101 表示仅需在第一个和第三个多义部件之间作出区分。“余”的 dist 项为 0，表明归约到“余”时多义已可消除,以后也用不着再做区分。

多义文法的控制算法如下:

1）在抽取基础部件时,若是多义部件,则置多义属性 *PO*.

2）文法归约过程中，将多义属性向父结点传递。若父结点的 EAG 定义中的 dist 项为 0，则将父结点的多义属性置 0，表明在这种情况下已不必区分。

3）当归约到根(整字)一级仍是多义结点,则以该结点的 EAG 定义中的 dist 项(此时非 0)为参数,沿归约成的树按多义属性向下寻找,直到多义基础部件。

4）启动该多义基础部件的多义区分函数，在 dist 项中值为 1 的位所表示的各选择项之间选择一个最可能的部件。

5）返回这一选择,即可获得多义区分的结果。

使用多义文法可以在提高畸变容纳程度的同时减少判断的工作量。例如，“日”是一

个大量使用的构字部件，但仅{日，曰}，{汨，汩}两对字需要在“日”与“曰”之间作出区分。此外多义区分函数具有很大的灵活性，它可以直接使用基础部件抽取算法，也可以使用专门的算法。由于区分是在指定的几个基础部件之间进行的，可靠性也可得到提高。

三、共 生 文 法

除了上述相似部件外，汉字集合中还存在着另外一种相似情况。例如“冈、同、月、冏、冂、周……”，尽管内部极不相象，但外轮廓却是相同的。人类识别这类结构往往只辨认其外轮廓就足够了，仅当存在极相似字的情况下才需要辨认内部。为了处理这一类情况，提出了共生文法的概念及其实现算法。

2-D EAG 中，定义共生等价类为

共生定义名：共生部件表；

例：对“冂”结构定义

冂：冈，同，月，…；

在如下的 EAG 定义中，

南，*UD*，十，冂，0；

钢，*LR*，钅，冈，111000；

“南”的 dist 项为 0，表明识别出“冂”已足以确认，不必再去识别“冂”的内部，而“钢”的 dist 项表明要在第一、第二和第三个共生部件之间作出区分，即判断到底是“钢”、“铜”还是“钥”。由于各种部件可在汉字集合中出现的部位并不一样，当同类结构太多时，还可以按照出现部位进一步细分为几个共生等价类。

共生文法控制算法的步骤如下：

1）在抽取基础部件时，若是共生部件，则建立共生文法结点，置共生属性 *CO*。共生结点表示的不是一个部件，而是由共生定义名代表的一个部件集合。

2）文法归约过程中，对共生结点中的每一个共生部件寻找文法匹配，只有匹配成功才将其父部件记录在父共生结点中。

3）在归约过程中，若共生结点中只剩下唯一一个共生部件并且相应的 EAG 定义中的 dist 项为 0，则将共生结点改造为普通的文法结点，并将共生属性置 0。

4）当归约到根(整字)一级仍是共生结点，则累计该共生结点中所有共生部件的相应 EAG 定义中的 dist 项，以此作为参数，沿归约成的树按共生属性向下寻找，直到终结符级的共生定义名结点。

5）启动该共生定义名的共生区分函数，在参数项中值为 1 的位所表示的共生基础部件间选择一个最可能的部件。

6）返回这一选择，即可获得共生区分的结果。

在上述算法中，步骤 3)消除共生的条件除了集合元素为 1 外，还要求 dist 项为 0，这使得共生文法所得到的结果与识别字集无关。dist 项可以根据所有汉字的集合设定，这样，即使某个汉字不在识别字集中，也不会误认为别的字。步骤 4）累计(求并运算)所有 dist 项是为了一次对所有需要区分的共生部件作出区分，以免反复调用区分函数。

共生文法体现了识别过程先外后内的策略。例如，“钢”字，按一般文法归约的顺序，应先归约钅、冈，再归约到“钢”，但在 2-D EAG 中，是先归约“钅”，“冂”，再归约到“钢”。当需要细分时，再按由上向下有引导的方式识别内部。

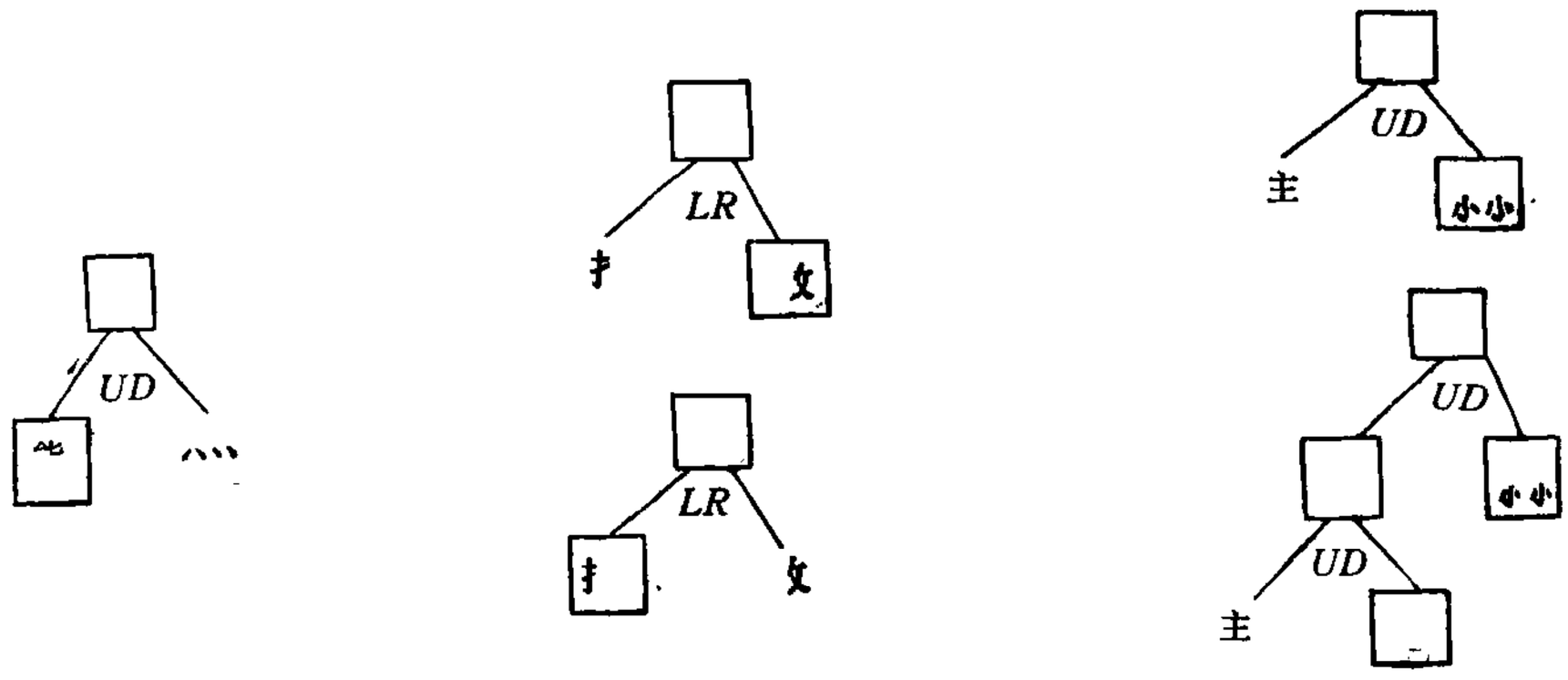
共生文法主要用来描述各种包围结构的部件(包围结构的汉字因一般需要区分内部，以另外的方法处理)。共生部件各自仍是完全独立的，与一般部件的定义完全一样，通过共生定义名定义一个共生等价类把一组共生部件联系到一起。归约时一组共生部件的集合作为一个文法结点，在归约的过程中，删除不可匹配的共生部件。若最后不必对共生等价类进行细分，就可唯一得到识别结果。当需要细分时，则启动该共生等价类的共生区分函数。

四、结 构 推 断

对结构简单的汉字，从四个方位(左上，右上，左下，右下)取基础部件后经不多于两级归约已能够识别。但对于结构复杂的字，这样做则还不够。例如，对“输”字，从四个方位只能取出“车、人、刂”。由于“人、刂”无法归约，故按常规归约方法不能识别。虽然把基础部件取大些(如把刖做为基础部件)就可以识别，但从实现效率来考虑，不能只顾少数复杂汉字而一味增大基础部件。此外，可以考虑抽取内部部件，增加归约层次的办法。但由于存在着粘连，直接抽取内部部件是不可靠的。为此，2-D EAG 中提出了结构推断的方法，用于识别结构复杂的汉字。

所谓结构推断，就是在利用已取出的部件无法判断到底是什么字的情况下，根据这些部件推测待识字的结构，建立推断树。然后按照推断树的结构，删除不可匹配的部件。若还需进一步区分极相似字的内部，则利用继承属性引导进一步的部件抽取，直到能够完成识别。

结构推断的关键是推断树的建立。最简单的情况是半字部件（汉字分解的第一级部件)可唯一确定。如“熊”，由于出现在下半字的“灬”必为半字部件，故可唯一建立一棵推断树。有些情况下半字部件不能唯一确定，如三分结构的“撤”，但可以知道“扌，攵”中必有

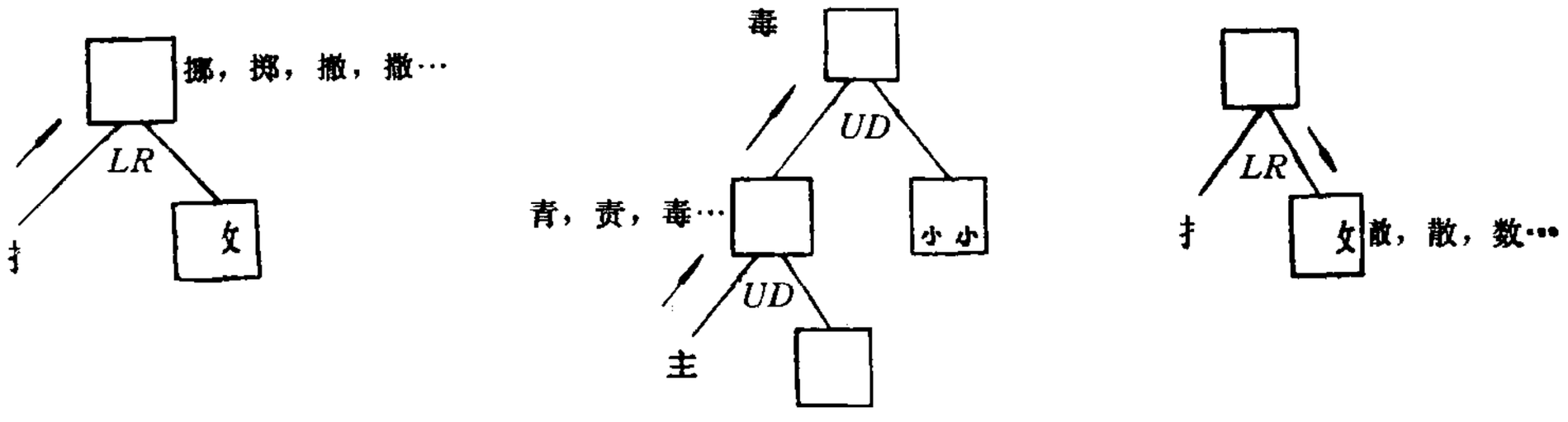


(a) 候选唯一确定　　(b) 有多个后选　　(c) 包括了二级推断树

图 1　结构推断树

一个为半字部件。这时可建立两棵候选树。还有一些情况下无法断定已取出部件中是否有半字部件，如“主”，它用于构成“青”时是半字部件，用于构成“㬎”时则不是。对于不能断定存在半字部件的情况，推断树的建立要复杂一些。

在图 1 中，没有被框起来的结点是索引部件，方框结点表示待推断部件，方框内的部件是已取出的基础部件。



(*a*) 按索引部件检索 EAG 定义　　(*b*) 二级推断树的检索　　(*c*) 与已取出部件匹配

图 2　推断树的检索匹配

结构推断算法如下：

1) 在已取出和已归约的部件中寻找半字部件。

a) 若可以唯一确定，建立唯一推断树(图 1——(*a*))。

b) 若可断定存在半字部件但不可唯一确定，则建立多个候选推断树 (图 1——(*b*))。

c) 若不可断定是否存在半字部件，则建立包括二级推断树的多个候选推断树 (图 1——(*c*))。

2) 以索引部件为关键词，索引 EAG 属性库，找出所有符合推断树分枝结构的 EAG 定义(图 2——(*a*))。

3) 若是二级推断树，则以这些 EAG 定义中的 CODE 项作为关键词，再一次检索 EAG 属性库，找出所有符合高一层推断树分枝结构的 EAG 定义(图2——(b))。

4) 把这些 EAG 定义中的另一部件作为待推断部件，寻找与已取出基础部件的匹配 (图 2——(*c*))。这是一个自上至下的匹配过程，待推断部件的字框相对于索引部件确定 (对二级推断树，还需要参照其他部件的位置)，由继承属性变量传递有关参数。

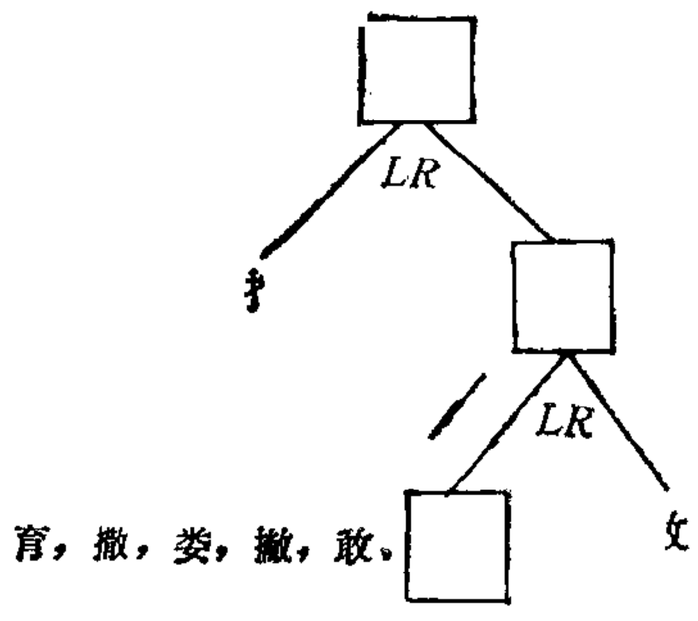


图 3　在待推断部件的引导下抽取内部部件

5) 若检索结果表明还不能够作出最后区分，则在候选待推推断部件的 EAG 定义的引导下，对字框内还没有识别出部件的部位进一步抽取部件，最后得到识别结果 (图 3)。

结构推断得到的识别结果的方式可以是全符合 (即EAG 定义中规定的每一个特征结构都要匹配)，使识别结果与识别字集的大小无关；也可以是唯一确认(当待推断部件唯一时即把它作为识别结果)，与识别字集的大小有关。

使用结构推断的方法还可以克服内部畸变的干扰。对“输”字，即使中间的“月”模糊

不清，由于根据已有部件做推断已可唯一确认，因而就用不着去辨认月了。

五、讨　论

2-D EAG 方法在扩展属性文法的基础上提供了多义文法、共生文法和结构推断这三种识别控制机制，它们利用汉字结构之间的相似性比较来抑制冗余识别，从而最大限度地减少了对汉字特征结构的不必要的依赖，提高了畸变容纳能力。

结构推断的方法除用于识别结构复杂汉字之外，也有可能用于识别特征结构缺损的汉字。如“朝”，因左上角“十”结构缺损而不能识别。但若进入结构推断，则将会得到结构推断树 

。可能符合这一树分枝的汉字有“朝，朔，期”。在三个部件“卓，𠦝，其”中择一就相对容易得多。这种做法非常类似于人类辨认结构缺损字的方式。要实现这一设想，首先要保证不能把结构缺损字误识作其他字，其次还需要有一个更好的部件完好度评价函数，以及一个更加灵活的部件处理顺序调度算法。

上述三种识别控制算法中都可以采取一些加速措施。在抽取多义基础部件时，若特征结构可唯一确认，如对部件“人”，则可以作一已区分标记。最后识别需要细分时就不必再启动区分函数，直接利用此标记即可。共生文法中，可以按不同的出现部位将共生等价类分，从而可减少共生结点中的共生等价名个数。而对于结构推断，则只须检索那些结构复杂的汉字(若不考虑结构缺损字的识别)，并且二级推断树的第一次检索只需考虑那些可用作字根的汉字或部件。

本文在董韫美教授的指导下完成，在此致以谢意。

参　考　文　献

[1] 赵　明，用于手写印刷体汉字识别的二维扩展属性文法方法，中文信息学报，1(1988)，3，78—84.
[2] Watt, D. A., Madson, O. L., Extended Attribute Grammars, *The Computer Journal*, **26** (1983), 2, 142—153.

THREE KINDS OF CONTROL STRATEGIES IN TWO-DIMENSIONAL EXTENDED ATTRIBUTE GRAMMARS

ZHAO MING

(*Software Institute, Academia Sinica P. O. Box* 8718, *Beijing* 100080)

ABSTRACT

This paper presents three kinds of control strategies, the polysemous grammars, the coexisting grammars, and the structure inferrings, realized in two-dimensional extended attribute grammar (2-D EAG) method for the recognition of hand written Chinese characters. With these control strategies, 2-D EAG method has higher tolerance ability to structural distortion. It makes analogy among similar structures, adopts the manner of “outline preference”, and constraints redundant recognition.

Key words: Chinese character recognition; two-dimensional extended attribute grammars; character structure analogies.