集成式故事语句分析模型 ISAM

中国科学院软件研究所 李小滨
100080 北京

摘要：本文探讨了故事理解的基础工作 —— 故事语义的获得，提出了一个集分词、语法和语义分析为一体的故事语句分析的黑板模型 ISAM，并且对一种新的词典组织方法 —— 字词网方法进行了讨论.

ISAM: An Integrated Analysis Model for Processing Story Sentences

Li Xiaobin
Institute of Software, Academia Sinica
100080 Beijing

Abstract: In this paper, an integrated analysis model for processing story sentences is presented and a new method for organizing lexicon is discussed.

一、引言

语句分析，不仅是故事理解的基础工作而且是自然语言处理最基本的内容。尽管这项研究在自然语言处理诞生之际就开始了，但至今远未达到理想的境界。

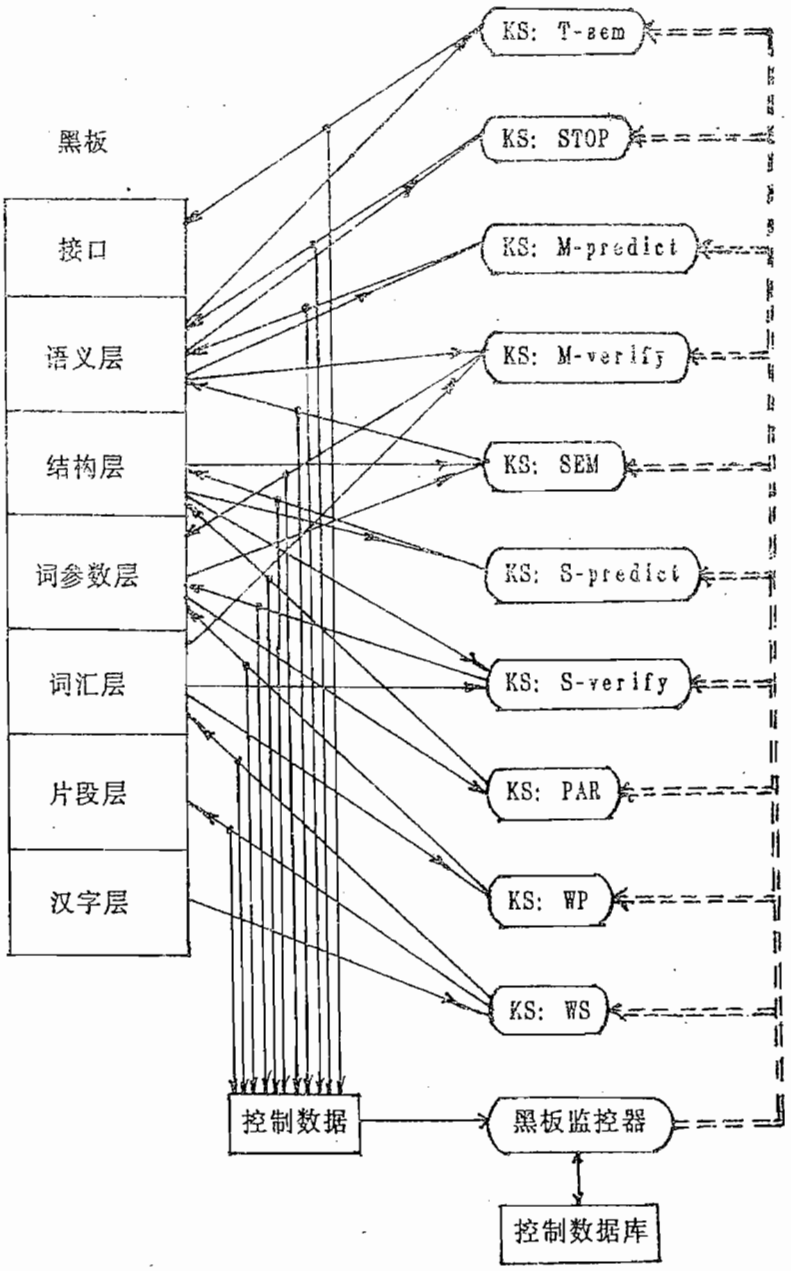
对于汉语故事语句的分析来说，传统的做法是串行地进行分析。即先对语句进行分词处理，得到语句的一个词序列；然后对这个词序列进行语法分析，得到语句的句法结构表示；最后对这个句法结构表示进行语义分析，得到语句的语义表示。然而这种串行的分析方式往往并不成功，在语句的这三方面分析中都可能出现棘手的二义性问题，如分词处理中的词链问题、语法分析中的二义性问题，以及语义分析中的歧义问题。尽管在过去的几十年里，人们在试图于本层面解决这些问题上付出了极大的努力，可是收效甚微。因为观察人类的语句分析行为可以注意到：语句的这三方面分析之间的依赖关系并不仅仅是单向的。即不仅仅是语法分析要依赖分词处理，语义分析要依赖语法分析，而且分词处理也要反过来借助局部语法及语义分析的结果，所以二义性问题不可能在本分析层面上孤立地得以解决。

如何利用这三方面分析之间的依赖关系来解决二义性问题呢？ 最初人们采用的办法是将部分语法、语义知识嵌入分词知识，将部分语义知识嵌入语法知识。但这样往往从整体上破坏了系统原有的模块性并且效果也不甚理想。近年来在语句分析上出现了将这三方面分析集成起来进行的趋势（如黄祥喜建立在词法 ATN 和句法语义 ATN 基础之上的分词理解并行处理系统，见[4]）。 这种集成式的语句分析不同于 Schank 早在 70 年代就提倡的那种集成式分析方法。Schank的那种集成分析不考虑分词处理并且是以语义分析为核心的，语法分析在其中只处于一个很次要的地位。而这里指的是分词处理、语法分析、语义分析三方面并重的集成式分析。下面提出的正是这样的一个故事语句集成式分析模型ISAM。

二、 ISAM 概述

语句集成分析模型 ISAM 是一个语句分析的黑板模型。语句分析要用到的所有知识被划分为若干独立的知识源（KS），每一个知识源完成一项特定的任务。如分词知识源 WS 进行分词处理；语法分析知识源 PAR 进行语法分析；语义分析知识源 SEM 进行语义分析；预测知识源 S-predict,M-predict 对后续词作出预测；验证知识源 S-verify, M-verify对预测信息进行验证；……等等。 语句分析的状态数据存放在一个黑板结构中，此黑板上的信息可被所有 KS 共享，KS 之间的通讯和交互只通过黑板进行。黑板监检器根据黑板的当前状态动态适时地选择和激活 KS， 使模型可在每个时刻对语句进行最适当的分析。ISAM 的结构如图 1 所示。

由 ISAM 的结构示意图可以看到： ISAM 主要由三部分构成：知识源、黑板结构及黑板监控器。下面将先介绍 ISAM 的词典、语法 ATN 及语义范畴体系，然后再分别对知识源、 黑板结构及黑板监控进行讨论。



注：“ ——→ ” 表示信息流　　“ =====> ” 表示控制流

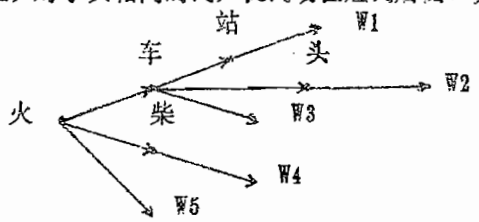
图　1 1 SAM 的结构

三、 ISAM 的词典及语法语义知识

上节已提到，在 ISAM 中语句分析要用到的所有知识被划分为若干相互独立的知识源。这里知识源之间的"相互独立"是指一知识源的执行不会对其它知识源产生任何直接的影响，而不是指各知识源之间涉及的背景知识不发生重叠。事实上，ISAM 的词典、语法知识及语义知识每一类均分属于多个不同的知识源，因此在讨论 ISAM 的知识源之前有必要先介绍一下它们。

1）ISAM 的词典

传统的机器词典类似于普通词典，其存贮上的冗余度是很大的，即由于一个汉字可以是许多词的构词成分，因此往往在词典中多次出现，为了解决冗余问题，有人试图在词典中以汉字树的形式来组织词汇（见[4]），即以某个字打头的所有词对应一棵汉字树；每个构词字均作为树的一个节点；含有相同构词字的词共用一条路径；对字头相同的词，长词动在短词后面。如：



显然汉字树的方法比原来传统词典的组织方法冗余度要小，但是冗余现象仍然是严重的，如"火车"和"车辆"，虽然它们的构词字都有"车"字，但是按汉字树的方法，词典中就无法只存放一个"车"字而让这两个词共享。为了进一步克服这种存贮上的冗余度并便于KS检索，我们提出一种字词网的方法来组织词典。即将词典组织成一种网状的结构（字词网）。网上的结点代表汉字，结点之间的有向弧则代表两结点上的汉字构成的词及其语法和语义特性。这样就解决了词典中数量最多的汉语双字词的表示问题。至于单字词，我们则在字词网上增设一个不与任何汉字相对应的"虚字"结点，一个字结点和"虚字"结点之间的弧则代表该字单独构成的词及其语法和语义特性。三个字以上构成的多字词在汉语中数量较少，这里采用的是在双字词表示的基础上将多字词的中间构词字增附在弧上的办法来解决。当然这时在弧上增附中间构词字会引起词典的冗余现象，但由于中间构词字数量很少，这种冗余现象较以往词典的冗余现象轻微得多。例子见图2。

下面给出字词网 CWN 的定义：

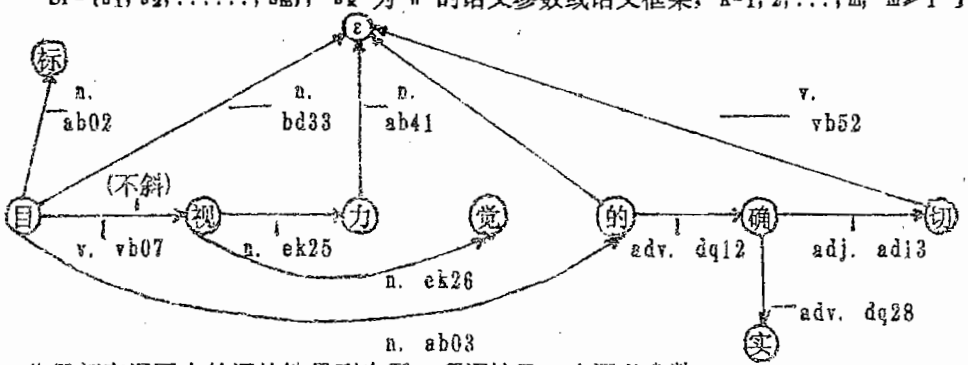
CWN = < V, E >

V = C ∪ N

C = $\{c_1, c_2, \cdots\cdots c_n\}$ 是词典中所有构词字的非空有穷集合。

N = { ε } 是仅含一个不与任何构词字相对应的"虚字"的集合。

E = { $(\langle v_i, v_j\rangle, M, (GP, SP))$ | $v_i \in C$, $v_j \in V$, $M = \varepsilon$ 或 $M = c_1' c_2' \cdots\cdots c_k'$,
$c_1', c_2', \cdots\cdots, c_k' \in C$, 且 $W = v_i M v_j$ 为词典中的词,
$GP = \{g_1, g_2, \cdots\cdots, g_l\}$, g_k 为 W 的语法参数, $k = 1, 2, \cdots\cdots l, l \geq 1$.
$SP = \{s_1, s_2, \cdots\cdots, s_m\}$, s_k 为 W 的语义参数或语义框架, $k = 1, 2, \ldots, m$, $m \geq 1$ }



注：此局部字词网中的词特性只列出了一项词性及一个语义参数。

图2 字词网局部示意

ISAM 中的词典按上述字词网的形式表示有以下几个特点：

a）冗余度小。除了多字词的少数中间构词字外，字词网上字不再出现重叠。

b）信息丰富。此字词网不仅包含了分词的信息，还包含关于词性、词义等信息。

c）易于检索和使用。如分词处理的 KS 很容易通过检索字词网而发现词链现象。当它检索发现输入语句的字序列中有一个子字序列（字数＞3）在字词网上的结点是按有向弧顺序相连的，便可确认此子字序列是一词链，否则它可容易地从此字序列中切分出词来。

2）语法知识

ISAM 中的语法知识以语法 ATN 的形式组织存放。由于语法 ATN 已广泛运用，这里就不再赘述.

3）语义知识

ISAM 中的语义知识除了包括在字词网中关于词的语义参数 sep 和语义框架 sef（对应于动词）外，还包括了一个语义范畴体系 SC。

$SC::=\{ssc_1, ssc_2, \cdots\cdots, ssc_m\}$

$ssc_k::=$ SC | sef | sep　　$k=1, 2, \cdots, m$

这里，$sef::=(sep, (ca_1, prop_1), (ca_2, prop_2)\cdots\cdots, (ca_n, prop_n))$

$ca_i::=$＜逻辑格名＞

$prop_i::=$＜语法约束＞＜语义约束＞

$i=1, 2, \ldots\ldots, n$

四、 ISAM 的知识源

ISAM 中语句分析的知识主要以知识源（KS）的形式表示，每个KS都对黑板进行操作， 完成某些特定的工作。KS 之间并不直接进行通讯和相互调用，而是通过黑板发生联系。一个 KS 可以通过建立或删除分析状态，修改或填充某一分析状态的属性值等方式向黑板添加信息。

每个 KS 都有下列成分：

Condition: 条件程序。一旦该 KS 被激活，进入激活 KS 队列 AKSQ（当黑板监控器根据黑板变化的类型及控制数据库中的信息确信该 KS 有可能执行时）， 此条件程序便详细检查黑板上的数据是否使该 KS 执行的条件成立。

Action: 动作程序。一旦条件程序确信该 KS 执行的条件成立时，此动作程序便可进入可执行的 KS 队列 EKSQ。在黑板监控器的调度下，它可被执行从而引起黑板上信息的变化。

在 ISAM 中共有 10 个 KS，每个 KS 涉及黑板中一个或几个信息层。下面简要地介绍各个 KS 的功能：

1）分词KS--WS

WS 以字词网 CWN 为背景从输入的语句字序列中切分出一个片段（当存在词链现象时）或一个词来。

2）词参数查寻KS--WP

WP 从字词网 CWN 中查找出词汇的语法参数（词性等等）及语义参数或语义框架（词义）.

3）语法分析KS--PAR

PAR 以语法 ATN 为背景，在词参数层提供的信息基础上对语句做局部或全局语法分析， 得到语句中短语或整个语句的句法结构树。

4）结构预测KS -- S-predict

S-predict 以语法 ATN 为背景由局部语法分析的结果对接在此结果前后的词汇作出预测。

5）结构预测验证KS -- S-verify

S-verify 从词参数层、词汇层及片段层中寻找根据以验证S-predict 所作预测的正确性。

6）语义分析KS -- SEM

SEM 以语义范畴体系 SC 为背景，在词参数层和结构层提供的信息的基础上对语句进行局部或全局的语义分析，得到语句的扩充格框架语义表示。

针对故事理解的需要，我们对 Fillmore 的格框架进行了改进从而提出了故事的一种扩充格框架表示 amf:

```
amf::= (no-csmf)
no-csmf::=no; csmf | no; csmf; no-csmf
no::= <语句序号>
csmf::= (rel-ssmf)
rel-ssmf::= rel; ssmf | rel; ssmf; rel-ssmf
```

ssmf::=pred pmf
rel::=<复合句里各简单句的关系>
pmf::={ca-val}
pred::=<谓词语义参数>
ca-val::=ca:val | ca:val;ca-val
ca::=<逻辑格关系> | <语义范畴>
val::=<语义参数> | pmf | ssmf
注：csmf 为复合句语义框架；ssmf 为简单句语义格框架；pmf 为短语语义框架.

上述的故事扩充格框架表示与 Fillmore 的格框架表示相比，有以下改进：第一，它不直接利用语句中的词来充当谓词或格框架中的格元素，而是用表示词义的语义参数来充当。这种语义参数可看成是词的深层表示，就使得语句的语义表示摆脱了语言表层结构遣词不同的影响，意义相同而遣词不同的语句也可以得到相同的语义表示。同时通过语义范畴还可使各语义参数发生联系，从而沟通了意思相近、意思相反等语句的语义表示间的联系；第二，它扩充了 Fillmore 格框架里所用谓词的含义.在故事的扩充格框架表示中谓词不仅可以为动词，而且可为任何合适的关键词。同时格关系也可根据需要而增设；第三，它除了有与谓词对应的格框架外，还有不与谓词对应的短语框架、复合句框架及故事框架。短语框架用于反映短语中各成分间的逻辑关系，语句框架用于指示各分句之间的内在关系，故事框架则体现各语句在故事中的先后顺序关系。

7）语义预测KS -- M-predict

M-predict以语义范畴体系 SC 为背景由局部语义分析的结果对其前后的词汇作出预测.

8）语义预测验证KS -- M-verify

M-verify 从词参数层、词汇层及片段层中寻找根据以验证 M-predict 所作预测的正确性。

9）分析结束KS -- STOP

STOP 确认系统分析终止的条件并在语义层选一最好的结果作为输出。

10）接口处理KS -- T-sem

T-sem 接收输出的语句语义表示，建立起它和故事中其它语句之间的顺序关系。

五、 ISAM 的黑板结构

在 ISAM 中，黑板作为全局工作区主要有两个用途：描述语句分析的中间状态，及结果；负责各 KS 之间信息的传递。整个黑板被划分为6个各自有一组黑板元素的信息层，每一层用于描述关于语句分析的某一类信息：

- 汉字层：用于记录语句输入时接收到的汉字。其黑板元素为字典中的所有字。
- 片段层：用于描述词切分时的语句片段划分情况。其黑板元素为片段标记。
- 词汇层：用于记录分词处理划分出来的词。其黑板元素为词典中的所有词汇。
- 词参数层：用于记录词的语法参数及语义参数。如词性、词义等等。 其黑板元素为语法参数和语义参数。
- 结构层：用于记录语法分析之后得到的短语结构、子句结构及语句结构，其黑板元素为句法结构树.
- 语义层：用于记录语义分析之后得到的短语的语义表示、子句的语义表示及语句的语义示。 其黑板元素为扩充格框架。

黑板各信息层与知识源的对应关系见图1。

黑板的主要信息载体是语句的分析状态。黑板上所有的分析状态都用统一的属性/值结构表示。在黑板上对语句分析状态的划分有两种：一个是按 Level 来划分，另一个是按 Time来分。ISAM 的分析可看成是在每一个 Level 上找出正确的分析状态。Level 不同的语句分析状态可以相互依赖。 一旦在最高信息层上找到了正确的语句分析状态，语句分析工作即告结束。Level 相同的所有可能的分析状态的集合实际上就构成了ISAM 在该层进行问题求解的解空间。至于 Time 对分析状态的划分则另有一种意义：Time 相同的情况下，Time 相同（时间上重叠）的分析状态在 ISAM 中是竞争的状态，因为它们代表了这一信息层上不同的分析结果。

六、 黑板的监控

ISAM 的控制部分是黑板监控器，它的任务是：

1）根据黑板的变化情况及控制数据库中的信息将合适的 KS选入 AKSQ。

2）检查 AKSQ 中各 KS 的条件（即运行各 KS 的条件程序），将条件满足的 KS 加入 EKSQ。
3）按某种择优原则从 EKSQ 中选出一个 KS，运行其动作程序从而使黑板上的信息发生了变化。在黑板的监控中，除了黑板变化的信息起重要作用外，控制数据库中的信息及监控器所采用的择优原则的作用也不容忽视。

控制数据库中存放的是黑板变化类型与可触发的 KS 之间的对应关系。例如图1中所示的黑板信息层与 KS 之间的对应关系（一旦黑板的变化发生在哪些层上，以那些信息层上的信息为输入的KS 就可认为是可触发的 KS 而选入 AKSQ）。

至于择优原则的规定是为了保证模型串行运行并避免组合爆炸，否则由于调用一个KS 可引起黑板上多种变化，而多种变化又可激活多个 KS，系统运行的任一时刻 AKSQ 和EKSQ 中都可能有许多KS，而这些 KS 的执行将生成竞争的 KS。ISAM监控器的择优原则是：
1）在高信息层上运行的 KS 要比在低信息层上运行的 KS 优先执行。
2）在同一信息层上运行的 KS，其执行结果改变了高信息层状态的 KS 要比改变了低信息层状态的 KS 优先执行。
3）输入、输出层次均相同的 KS，时间上先进入 EKSQ 的 KS 优先执行。
4）利用的数据正确性越高的 KS 越优先执行。有时某些 KS执行后得到的结果并不绝对可靠，这些不绝对正确的数据因此就可能出现在黑板的某信息层上。显然这种不可靠的信息多次传播或多次叠加都将产生更加不可靠的信息。
5）任务越简单（执行代价小）且得出的结果越正确的 KS 越优先执行。

择优原则的 1）--3）体现了 ISAM 的某种语句分析策略，即它不象传统的语句分析那样总是先在低层进行尽可能多的分析后才转入高层的分析，例如输入若干个字它就开始分词，切出若干个词它就开始确定词的参数并进行语法分析，得出若干语句的局部结构又接着试图作局部的语义分析，等等。当然 ISAM 的这种策略决不是单纯的自底向上的分析，结构层和语义层上的信息变化都可导致预测 KS 的执行，从而转为自顶向下的分析。

择优原则4）体现了ISAM 的正确性原则。择优原则5）则体现了ISAM 的功效原则。这些择优原则的综合运用将使得黑板监控器能根据黑板的变化信息动态适时地执行某个 KS，体现 ISAM 对语句的机遇性分析。

致谢　本文得到导师陆汝钤研究员的指导，在此深表谢意。

参考文献

[1] 李小滨、徐越，“自动文摘系统 EAAS”，《软件学报》，Vol.4，1991.
[2] Li Xiaobin, IUC: An Interface for Understanding Chinese Language, Advance in Chinese Computer Science, Vol.2, 1989.
[3] H. Penny Nii，“黑板系统，”《计算机科学》，No.5- No.6，1987.
[4] 黄祥喜，“书面汉语的计算机分词和理解”，吉林大学博士论文，1989.
[5] James Allen, Natural Language Understanding, 1987.
[6] Avron Barr and Edward A. Feigenbaum, The Handbook of Artificial Intelligence, Vol. I, 1980.