

英文生成系统词汇结构与功能位的逻辑描述

卞世力 姚天顺 滕永林

(东北工学院计算机系 沈阳110006)

摘 要

This paper presents the structure and function of lexicon used to generate English from intermediate language in Chinese English Machine Translation (CETRAN). The transformation principles between each grammatical levels for the generation system are described. Some examples go through this paper to explain the method of English generation system. A grammar—Lexical Functional Position Grammar, which is applied to the generation system in the CETRAN, is advanced.

随着人工智能(AI)各领域的不断发展,越来越需要以自然语言形式进行人机对话,使其更加面向用户,特别是机器翻译的再度兴起,使得自然语言处理(NLP)面临新的挑战。在近几年,国内外一些语法理论陆续产生,如:管辖与约束理论(Government Binding Theory),词汇功能文法(Lexical Functional Grammar),功能合一文法(Functional Unification Grammar)和词汇语义驱动理论等,但是不管你是用哪一种方法,都离不开带有复杂特征集(Complex Feature Set—CFS)的词做为知识源。CETRAN系统生成器也是由CFS支持的,而且对于一个词X的复杂特征集有 $CFS(X) = (MNI(X), SFL(X), DFL(X))$ 。其中:MNI(X)—X主结点

信息;SFL(X)—静态属性表;DFL(X)—动态属性表。CFS(X)在这三个特征子集中又包含着许多特征信息,使得知识源十分丰富,系统的效率大大提高了。

CETRAN系统采用的是以中间语言媒介语做枢轴结构,力求这种枢轴结构能够成为独立于语种的通用性语言接口,目的是为此后汉英双向翻译,以及多语种翻译打下基础,整个系统不论是分析或生成系统都是基于词义的,利用CFS通过语义驱动自然语言的处理工作。

一、词汇结构与复杂特征集

CETRAN生成系统是分析系统的逆过程,是逐层次进行转换的,即从语义→句法→词法→输出目标语句。一个词X的CFS也

收到日期: 92-11-03.

因此去选择另外一个介词短语PP,修饰the chair,语义解释器发现the window是表明the chair的位置,因此,在连接时,by被指定为(LOC)关系。

4. 小结

概念图模型作为未来基于知识系统和自然语言处理系统的有效知识表示方法,它能够处理象时态、模态等问题,能灵活地表

示大量的语义词汇。几年来经过西北大学人工智能实验室全体同志和研究生认真的理论研究和“概念图处理器”工具系统的开发及英文的自然语言解释和汉语的自然语言生成的系统模拟^[5,6,7,8],证明了概念图模型是适合自然语言处理的知识表示方法。目前我们正在研究开发一个基于概念图的中英文翻译系统。(参考文献略)

可分为四层次：语义逻辑层 (SL)，句法层 (SY)，词法层 (MO)，音位层 (PH)。所以以一个X的复杂特集可表为：

$$CFS(X) = (SL(X), SY(X), MO(X), PH(X)) \dots (1)$$

$$CFS(X) = (MNI(X), SFL(X), DFL(X)) \dots (2)$$

(1) 是将特征按层次划分的，(2) 是按知识类型划分的。因此将(1)(2)统一起来，可以描述CFS结构为：

$$CFS(X) = (MNI(X) = \begin{matrix} SL(X) \\ SY(X) \\ MO(X) \\ PH(X) \end{matrix}, SFL(X) = \begin{matrix} SL(X) \\ SY(X) \\ MO(X) \\ PH(X) \end{matrix}, DFL(X) = \begin{matrix} SL(X) \\ SY(X) \\ MO(X) \\ PH(X) \end{matrix})$$

因此在我们系统中，一个词X的复杂特征集可以描述为：

$$CFS(X) = (MNI(X), SFL(X), DFL(X))$$

$$MNI(X) = \begin{matrix} SL(X) = (Semrela(X), \\ \quad centre(X), rest(X) \dots) \\ SY(X) = (cat(X), subcat(X), \\ \quad synrela(X) \dots) \\ MO(X) = (Lex(X), \\ \quad china(X) \dots) \\ PH(X) = SPL(X, \langle \text{音标/拼音} \rangle) \\ SL(X) = (rest(X), syno(X), \\ \quad match(X) \dots) \end{matrix}$$

$$SFL(X) = \begin{matrix} SY(X) = (cat(X), subcat(X), \\ \quad syntea(X), patt(X) \dots) \\ MO(X) = (wdph(X), \\ \quad form(X) \dots) \\ PH(X) = spl(X) \\ SL(X) = (dmch(X), loc(X), \\ \quad subfea(X) \dots) \end{matrix}$$

$$DFL(X) = \begin{matrix} SY(X) = (mood(X), \\ \quad voice(X), dtime(X) \dots) \\ MO(X) = (phrase(X) \dots) \end{matrix}$$

例如：英语 read 一词

$$CFS(read) = (MNI(read), SFL(read),$$

$$DFL(read))$$

$$MNI(read) = \begin{matrix} SL(read) = (SEN, *, 2036, \dots) \\ SY(read) = (v, vv2, vp, \dots) \\ MO(read) = (read, 看, \dots) \\ PH(read) = [ri:d] \end{matrix}$$

$$SFL(read) = \begin{matrix} SL(read) = 2036, sec, \\ \quad (\#1111, \#1311) \dots \\ SY(read) = (v, vv2, vp6a, \dots) \\ MO(read) = (w, \\ \quad 1/vs4/ed5/ig1, \dots) \\ PH(X) = [ri:d] \end{matrix}$$

$$DFL(read) = \begin{matrix} SL(read) = (want\ to, \dots) \\ SY(read) = pre, \dots \\ MO(read) = (, \dots) \end{matrix}$$

从MNI(read)中可知read是句子的谓语，中心词。从DFL(read)中可知read是现在时而且是want to read。从SFL(read)中可知read左为人类名词/人称代词(语义码为1111)，右边为书写物名词(语义码为1311)等。

二、语义逻辑层的逻辑描述

在CETRAN中，分析系统是将汉语句子转换成中间媒介语，我们系统的中间媒介语语言实质上是一个由语义关系图 $M = \langle V(M), R(M), \Psi_m \rangle$ 表示的。其中： $V(M)$ ——是一个非空结点集合，每个结点是一个带有复杂特征集的概念； $R(M)$ ——是边集合，每个边表示一个语义关系(共有50种关系)； $\Psi_m(R)$ ——是从边集合R到结点的有序偶集合上的函数。如图1：

SENTENCE

$$\begin{matrix} | \\ \text{---SEN} \rightarrow \text{watch} : [CFS(\text{watch})] \\ \\ \text{---EXT} \rightarrow \text{television} : [CFS(\text{tele-} \\ \quad \text{vision})] \\ \\ \text{---AGT} \rightarrow \text{he} : [CFS(\text{he})] \end{matrix}$$

图1 “他在看电视”中间语言

其中： $V(M) = \{\text{watch}, \text{television}, \text{he}, \text{SENTENCE}\}$ ； $R(M) = \{\text{SEN}, \text{EXT}, \text{AGT}\}$ ； $\Psi_m(\text{SEN}) = (\text{SENTENCE}, \text{watch})$ ； $\Psi_m(\text{EXT}) = (\text{watch}, \text{television})$ ； $\Psi_m(\text{AGT})$

$\Rightarrow (\text{watch}, \text{he})$ 。

说明:SENTECNE是句类标识结点;SEN是句类标识关系,不属于50种语义关系。

在图1中,CFS(X)中只有MNI(X)和DFL(X)并没有SFL(X),需要用MNI(X)中的rest(X)和cat(X)与英语词典中的erest(X)和ecat(X)去匹配,若匹配成功则该词被选中,取该词的SFL(X),否则,去匹配SFL(X)中提供同义词syno(X),直到匹配成功,最后得到完整CFS(X)。逻辑描述为:

TEST((rest(X), erest(X)), (cat(X), ecat(X))) $\Rightarrow$ TAKE(SFL(X))

若匹配成功则该词被选中,取该词的SFL(X)。

-TEST((rest(X), erest(X)), (cat(X), ecat(X))) $\Rightarrow$

TEST((rest(syno(X)), erest(syno(X))), (cat(syno(X), ecat(syno(X))))。

否则,去匹配SFL(X)中提供的同义词syno(X),直到匹配成功。

在得到完整的CFS(X)后,根据语义逻辑关系SLR将中间语言的树形转换成线性的偏序集 $\langle S, \leq \rangle$,S是R的子集,其中S一为该句的词集合;R一为该语种词的全集; $\leq$ 一为序关系。

若设图1的各元素为 $X=\text{watch}$, $Y=\text{television}$, $Z=\text{he}$, 则该图可用 ψ_m 描述语义关系为:

$\psi_m(\text{SEN}) = (\text{SENTENCE}, X)$ 。

$\psi_m(\text{EXT}) = (X, Y)$ 。

$\psi_m(\text{AGT}) = (X, Z)$ 。

因此将图1, 转换 $\leq$ 关系的逻辑描述:

$\psi_m(\text{SEN}) = (\text{SENTENCE}, X) \Rightarrow$

$\text{SENTENCE} \leq X$

$\psi_m(\text{EXT}) = (X, Y) \wedge \text{ecat}(X) = v \Rightarrow X \leq Y$

$\psi_m(\text{AGT}) = (X, Z) \wedge \text{ecat}(X) = v \Rightarrow Z \leq X$

得到一个线性序关系集:

$\text{LRS}(S) = \{X, Y, Z | \text{SENTENCE} \leq X, X \leq Y, Z \leq X\}$ 。

如何确定这些元素的全序关系, 可以根据下面的 $\leq$ 关系的两个性质进行确定: 设 $\alpha \in S$, $\beta \in S$, $\gamma \in S$

性质1: $\alpha \leq \beta \wedge \beta \leq \gamma \Leftrightarrow \alpha \leq \beta \leq \gamma$

性质2: $\alpha \leq \beta \wedge \gamma \leq \beta \Leftrightarrow \alpha \leq \gamma \leq \beta$

$\gamma \leq \beta \wedge \alpha \leq \beta \Leftrightarrow \gamma \leq \alpha \leq \beta$

$\text{LRS}(S) = \{X, Y, Z | \text{SENTENCE} \leq X, X \leq Y, Z \leq X\}$

性质1

$\Leftrightarrow \{X, Y, Z | \text{SENTENCE} \leq X \leq Y, Z \leq X\}$

性质2

$\Leftrightarrow \{X, Y, Z | \text{SENTENCE} \leq Z \leq X \leq Y\}$

得到一个线性链表:

$\text{SENTENCE} \rightarrow Z \rightarrow X \rightarrow Y$
 $\{ \text{CFS}(X) \} \{ \text{CFS}(Y) \} \{ \text{CFS}(Z) \}$

三、句法层的功能位描述

任何一种语言表达正确的意义, 即要输出一个正确句子S, 需要确定两部分内容。一部分是组成这个句子的各个词项间的顺序, 即词项在句中的位置; 另一部分是确定每个词项应具有形态。我们的系统是根据分析系统提供的中间语言意义来确定每个词项X应在句中的位置Position(X)和应有的词形form(X), 生成一个表达该意义的目标语的句子。

例如: 在图1中, 句子有三个位置(可命名为: 主位=pos1, 心位 pos2, 宾位 pos3)

$\begin{array}{|c|c|c|} \hline \text{pos1} & \text{pos2} & \text{pos3} \\ \hline \end{array} \rightarrow \text{position}(X)$

对于 watch 一词则有三个位置可选。根据图1的语意关系可确定 position(watch)=pos2 而 watch 的形态可选项又为:

watch——nos (原形)
 watched——ed1 (过去式)
 watched——ed2 (过去分词)
 watching——ing (现在分词)
 watches——s (单第三人称现在时)

 form(watch)= 另外还有由于时、态、句型引起的加助动词的形态

 do watch——do+nos (疑问句)
 do'nt watch——do+nos (否定句)
 will watch——will+nos (将来时)

 is watching——is+ing (现在进行时态)

同样根据图1的语义关系和每个词的DFL(X), 英语语法规则, 在纵向上确定了form(watch)=is+ing。

我们称一词项X在一个句中的这种由横向position(X)和纵向form(X)所确定的位置, 为该词项X的功能位, 记为FP(X), 且Position(X)记作P(X), form(X)记作F(X)。一词X的功能位可写成FP(X)=(P(X), F(X)), 如图2所示。

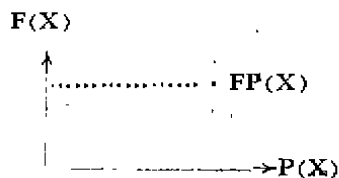


图2 功能位FP(X)示意图

3.1 确定P(X)

P(X), 我们是用偏序关系 $\leq$ 来确定的。如图1的watch可描述成: $P(\text{watch}) = \{\text{watch} | \text{SENTENCE} \leq \text{watch}, \text{watch} \leq \text{television}\}$ 或 $P(\text{watch}) = \{\text{watch} | \text{SENTENCE} \leq \text{watch} \leq \text{television}\}$ 。

3.2 确定F(X)

根据动词的时、态, 名词的单、复数及句型(陈述, 疑问句...) 确定F(X)值。如:

若X的DFL(X)中的dtime=pre(现在时), voice(X)=prg(进行态), 则form(X)=is+ing。

逻辑描述: $(\text{dtime}(X) = \text{pre}) \wedge (\text{voice}(X) = \text{prg}) \wedge (\text{cat}(X) = v) \Rightarrow \text{form}(X) = \text{is} + \text{ing}$ 。从而可得图1中各节点的功能位:

$FP(\text{he}) = (P(\text{he}), F(\text{he}))$
 $P(\text{he}) = \{\text{he} | \text{SENTENCE} \leq \text{he}, \text{he} \leq \text{watch}\}$
 $F(\text{he}) = \text{nos}$
 $FP(\text{watch}) = (P(\text{watch}), F(\text{watch}))$
 $P(\text{watch}) = \{\text{watch} | \text{he} \leq \text{watch}, \text{watch} \leq \text{television}\}$
 $F(\text{watch}) = \text{is} + \text{ing}$
 $FP(\text{television}) = (P(\text{television}), F(\text{television}))$
 $P(\text{television}) = \{\text{television} | \text{watch} \leq \text{television}, \text{television} \leq \text{END}\}$
 $F(\text{television}) = \text{the} + \text{nos}$
 在机器内部实际是一个键表:
 SENTENCE, (, SENTENCE)
 he, (W_FORM=(NULL))
 is, (W_FORM=(NULL))
 watch, (W_FORM=(add_ing))
 the, (W_FORM=(NULL))
 television, (W_FORM=(NULL))
 END, (, END)

在词法层中, 可根据W_FORM的特征值对每个词进行形态转换, 得到一个完整的句子: He is watching the television.

结束语

利用语义网络描述中间语言, 生成一个与其同义的句, 并不是一件很容易的事, 其中包括一些本文未介绍的调序, 增加节点, 删除节点以及词性转换等处理。本系统正象本文介绍的, 着重强调了词汇功能, 利用语义逐层转换的, 所提出思想和方法已在5万多词的词典环境下做过几百句试验, 证明是可行的。随着问题不断深入, 肯定还会有不少问题, 请批评指正。

74-77

计算机科学1993 Vol.20 No.5

计算机

声音处理

15

计算机声音处理技术

TP391

胡上序 钱涛

(浙江大学计算与信息中心 杭州310027)

摘 要

Dealing with voice in computer is becoming an active and important field of computer science. This paper summarizes various kinds of available technologies of handling voice in computer and shows the latest development of this field. Moreover, the speech recognition technology is discussed in details.

一、引言

计算机从处理单一的文字和数据发展到能够处理声音是一个质变。早先计算机通讯多数是采用发送电子邮件或通过网路系统共享数据和文本文件。仅仅限于这种方法是远远不够的。如同人类进行交流的方式除了书写成文的东西,还有语言。计算机声音处理因此也是计算机研究领域的一个重要方向。该技术涉及到诸多其它领域的技术,如电子学、通讯技术和控制学。目前这方面的研究取得了较大的进展,技术也日趋成熟。

二、计算机声音处理技术

计算机声音处理技术可以分为三大类:

信号内容处理、连接控制和软件构造。

2.1 信号内容处理

信号内容处理是指创建、操纵和分析音频电路中的信息。这种技术主要包括音频信号数字化、数字信号处理、文字-语言的合成和语言识别等技术。其中,音频信号的数字化编码技术是目前较为成熟的技术。

2.1.1 音频信号数字化 将模拟音频信号数字化是现代化电话系统中的一个最显著的特征。大多数现代化交换机设备都是将音频电路信号转化为数字串流的。

标准电话系统使用3字节脉冲编码调制(PCM),并产生传播速率为64k/s的数据

胡上序 教授,博士导师,主要从事语言识别,计算机神经网络,计算机仿真与智能信息系统等研究,钱涛 博士生,主要从事多媒体技术,计算机语言识别等方面的理论研究和实际系统开发。

参考文献

- [1] Bian Shili, Yao Tian Shun, A system of generating target language from interlingua, 5th SSICCA., 1992
- [2] R. Simons and J. Slocum, Generating English Discourse from Semantic Networks, Communication of the ACM Vol 15 No 10
- [3] 左孝凌,刘永才等《离散数学》上海科学

技术出版社

- [4] 殷钟嵘等《英语语法理论及流派》四川大学出版社
- [5] 姚天顺、马黎环, Generating English Text from Chinese Semantic Representation, Proceedings of 1988. International Conference on Computer Processing of Chinese and Oriental Languages Montreal Canada 1988