

自然语言合理句子的生成系统

胡 玥^{1,2)} 高小宇²⁾ 李 莉¹⁾ 高庆狮^{1,2)}

¹⁾(北京科技大学信息工程学院 北京 100083)

²⁾(中国科学院计算技术研究所 北京 100190)

摘 要 基于语法规则的使用 Chomsky 上下文无关文法产生的句子虽合乎语法规则但并不一定合乎语义常理. 文中提出一种可以产生合理的自然语言句子的产生系统 $\{T_I, N, B_I, C_I, S, F_I, P_I\}$, 其中, I 是指定的自然语言. 该系统建立在语义语言理论上, 使用 Chomsky 短语语法. 系统的关键技术是引入了语义单元技术、类型参量, 类型参量的作用不仅体现在语义单元中而且还体现在句树、事物树和函数集中. 语义单元和类型参量共同作用能有效的对产生的句子进行语义约束使得生成的句子合理.

关键词 多自然语言; 生成系统; 语义单元; 语义语言; 多语言机器翻译

中图法分类号 TP18 DOI 号: 10.3724/SP.J.1016.2010.00535

A Generating System to Generate Legitimate and Reasonable Sentences

HU Yue^{1,2)} GAO Xiao-Yu²⁾ LI Li¹⁾ GAO Qing-Shi^{1,2)}

¹⁾(School of Information Engineering, University of Science & Technology Beijing, Beijing 100083)

²⁾(Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190)

Abstract The sentences produced by Chomsky's Context-free Grammar are legitimate, conform with syntax rules, but not reasonable. The sentences produced by $\{T_I, N, B_I, C_I, S, F_I, P_I\}$ producing system proposed in this paper are legitimate and reasonable, where I is a specify natural language. It is based on semantic unit theory and Phrase-structure Grammar. The key of reasonableness is based on semantic-units theory, to introduce the category-parameter system, which is in the set $B-R_I S U$, in the set $C-tree_I$, and in the set F_I . The semantic-units and category-parameters make the sentences reasonable.

Keywords multiple natural languages; generating system; semantic unit; semantic language; multiple languages machine translation

1 引 言

(1) N. Chomsky 形式文法理论
N. Chomsky^[1] 在 1956 年提出了形式文法理论.

该理论是理想的形式化的描述工具之一, 不仅对语言学, 而且对整个计算机科学起到了很重要的作用.

但是, 基于语法的 Chomsky 文法生成的句子虽合乎语法规则, 但不一定合乎语义常理. 其根本原因是生成规则只反映语法规则, 不能有效反映

收稿日期: 2009-05-06; 最终修改稿收到日期: 2009-12-28. 本课题得到国家“九七三”重点基础研究发展规划项目基金(2007CB311103)、国家“八六三”高技术研究发展计划项目基金(2006AA01Z140, 2006AA010101)、国家自然科学基金(60873002, 60736013)资助. 胡 玥, 女, 1963 年生, 博士, 副教授, 主要研究方向为自然语言处理、计算机体系结构和网络安全. E-mail: huhuyue_001@sina.com. 高小宇, 男, 1978 年生, 硕士, 主要研究方向为自然语言处理、网络安全和串匹配. 李 莉, 女, 1980 年生, 博士, 讲师, 主要研究方向为自然语言处理和串匹配. 高庆狮, 男, 1934 年生, 中国科学院院士, 主要研究领域为并行算法、计算机体系结构、自然语言处理、人类智能、网络安全和串匹配算法.

语义知识.

(2) J. J. Katz 的语义标式(semantics marker)

J. J. Katz 的语义标式^[2]相当于语素分析表达式,是建立在词-词组-句、词的义素(semanteme)和“句子的语义是句子中词义总和”的传统思想的基础上的,虽然有很好的递归原则(recursive principle)的思想,但是仍然无法实用.

(3) C. J. Fillmore 的格文法(case grammar)

C. J. Fillmore(菲尔墨)的格文法^[3]反映了句子各语义成分之间的本质语义关系.这些本质语义关系独立于具体的自然语言语种.

(4) 从动词的“配价”理论到“语义单元”理论

法国语言学家 L. Tesnière 在 20 世纪 50 年代末从结构句法(stratural syntax)的角度提出的以动词为句子中心的配价语法^[4],是语言学上的重要贡献.其中,动词的价是指该动词支配名词性成分的数目.而在 20 世纪 80 年代提出的带类型参量的语义单元理论^[5-11]具有以下特点:① 动作类型的语义单元的参量不限制为名词类型,可以是形容词、数词、副词、句子或者简化的句子等等类型;② 数目也不是被限制在 3 个或者 3 个以下,可以是 4 个或者超过 4 个.③ 尤其重要的是,一个或者多个实量(每个实量是指一个或者多个字串)和一个或者多个虚量(每个虚量是指一个或者多个类型参量之“或”)是融合为一体的,不可分隔.也就是,不仅限于动词与其相关的名词(及其他词)之间的语法上的依存关系,而是全部实量、虚量融成为一体,形成一个独立语义的语言单位,称为该自然语言上的“语义单元表示”.④ 该理论中,实量不仅限于动词,还包括相关的词(例如,介词、冠词、副词、助词,等等).⑤ 句子就是“语义单元表示”中的类型参量不断地被该类型的其他语义单元表示所代入,直到全部都是实量.⑥ “语义单元表示”的语义就是意义单位(unit of meaning),称为“语义单元”(semantic unit).⑦ 更重要的是:语义单元与它在各个自然语言(例如,英、汉、日、等等)的语义单元标准表示 representation 是一一对应的.⑧ 而且它在各个自然语言的表示中,实量的个数和实量、虚量的位置(place)是可以各不相同的.例如,一个语义单元的英语表示“〈Noun₁〉 consists of 〈Noun₂〉”,其对应的汉语的语义单元表示是“〈Noun₁〉是由〈Noun₂〉组成”.

我们首先在第 2 节讨论自然语言合法且合理的统一生成系统,所涉及有关的内容将分别在其后的第 3 节到第 8 节中讨论.

2 自然语言合法且合理的句子生成系统 LRSGS

生成自然语言的合法而且合理句子的生成系统 LRSGS 是建立在语义单元及其表示理论上的一个 0-型形式系统

$$LRSGS_I = \{T_I, N, B_I, C_I, S, F_I, P_I\}, I=1, 2, \dots,$$

其中,

T_I 是在 I -自然语言上的终结符集;

N 是独立于具体语言的类型非终结符集;

S 是独立于具体语言的初始符, $S \in N$.

$B_I = \{B-R_I S U\}$ 是 I -自然语言上的带类型(category)参量或者不带类型参量的基本语义单元表示集,其中,类型参量是非终结符或者非终结符的“逻辑或”,其它部分是 I -自然语言上的终结符串.参见第 3 节.

$C_I = \{C-tree_I\}$ 是 I -自然语言上的句树与事物树,通称为类型树.在 I -自然语言上类型树上的所有的叶节点是 I -自然语言上的基本语义单元表示 $B-R_I S U$;所有在 I -自然语言上的类型-树的非叶节点是由基本语义单元表示的类型所组成的类型节点,对所有的 I -自然语言都一样.参见第 4~6 节.

F_I 是 I -自然语言上的函数集 $\{F_I(X_I)\}$,其中, $X_I, F_I(X) \in (T_I \cup N)^*$,即集合 $T_I \cup N$ 上的字串.参见第 6 节.

P_I 是 I -自然语言上的生成式集 $\{\alpha_I \rightarrow \beta_I\}$,其中, $\alpha_I, \beta_I \in (T_I \cup N)^*$,即集合 $T_I \cup N$ 上的字串.参见第 7 节.

I 表示任意一种自然语言,如 $I =$ 英语或 $I =$ 汉语.

3 语义单元与语义表达式

3.1 基本语义单元(B-SU)及其在 I -自然语言上的表示(B- $R_I S U$)

不同的 $K(K > 1)$ 种自然语言(例如,英语、汉语、日语、…)的句子之间之所以可相互翻译,是因为这些对应的句子之间有相同“意思”(即语义).

我们称句子的语义为“句义”,称句子为其句义在该自然语言- I 上的表示(即句义表示,记作 $R_I S S$).

在 K 个不同的自然语言中表示相同句义的句子中,表达相同“意思”(即语义)的单元称作“语义单元表示”(记作 $R_I S U$);该“意思”(即语义)被称作

“语义单元”(记作 SU)^[5-12].

显然句义是语义单元,句子是语义单元表示.

语义单元表示可以是:全部由在 I -自然语言上的终结符集 T_I 上的字(称为实量, real-parameter)组成,即简单语义单元表示(记作 $S-R_I SU$);也可以由终结符集 T_I 上的字(实量)和由非终结符集 N 元素表示的类型参量(type parameter, 称为虚量, pseudo-parameter)所组成,称为一般语义单元表示;还可以只由虚量所组成,称为抽象语义单元表示.

简单语义单元表示是不带类型变量的语义单元表示,一般语义单元表示和抽象语义单元表示是带类型变量的语义单元表示.

3.2 句义表达式

句义表达式(记作 $SS-exp$)就是用基本语义单元通过代入形成的句义.

I -自然语言上的句子表达式(记作 S_I-exp)就是用基本语义单元表示通过代入形成的句子. 显然,

S_I-exp 是 $R_I SU-exp$.

例如,句子 1:“Mr. Smith is an engineer”(英语) \leftrightarrow “史密斯先生是工程师”(汉语) \leftrightarrow “シミシさんほ技师です”(日语).

其在英语上的句子表达式 S_E-exp 是“‘Mr. ‘Smith’ is an ‘engineer’”;

其在汉语上的句子表达式 S_C-exp 是“‘史密斯先生’是‘工程师’”;

其在日语上的句子表达式 S_J-exp 是“‘シミシさん’ほ‘技师’です”.

其句义表达式是“ $1(2(4), 3)$ ”,即 $SS-exp = 1(2(4), 3)$.

其中,4 是 Smith 对应的语义单元;3 是 engineer 对应的语义单元;2(\langle 姓名|姓|名 \rangle)是“Mr. \langle 姓名|姓|名 \rangle ”对应的语义单元;1(\langle 表示人的名词 \rangle , \langle 表示具有职称的人的名词 \rangle)是“ \langle 表示人的名词 \rangle 是 \langle 表示具有职称的人的名词 \rangle ”对应的语义单元. 参见表 1.

表 1 语义单元及其在自然语言上的表示

基本语义单元	语义单元英语表示	语义单元汉语表示	语义单元日语表示
1(\langle 表示人的名词 \rangle , \langle 表示职称的名词 \rangle)	\langle 表示人的名词 \rangle is a \langle 表示职称的名词 \rangle	\langle 表示人的名词 \rangle 是 \langle 表示职称的名词 \rangle	\langle 表示人的名词 \rangle ほ \langle 表示职称的名词 \rangle です
2(\langle 姓名 姓 名 \rangle)	Mr. \langle 姓名 姓 名 \rangle	\langle 姓名 姓 名 \rangle 先生	\langle 姓名 姓 名 \rangle さん
3	Smith	史密斯	シミシ
4	engineer	工程师	技师

句义表达式及句子表达式可以通过语义单元表示树剪枝快速获得,具体可参考文献[7].

全部“语义单元”构成“语义语言”. 全部“语义单元在某个自然语言上的表示”构成该自然语言. 这里的“自然语言”是指英语、汉语、日语等等.

3.3 句子的构成

自然语言的句子不是由字或词线性排列所构成,而是由“语义单元表示”和“代入”运算所构成.

字母只是构成字的元素,它本身没有独立意思. 同样,字只是构成语言单元的元素,它本身同样没有独立意思.

例如,字“time”,它本身没有独立意思. 语义单元表示“ \langle 数 \rangle times \langle 数 \rangle ”是“ \langle 数 \rangle 乘 \langle 数 \rangle ”,而语义单元表示“ \langle 数 \rangle times”是“ \langle 数 \rangle 次”,语义单元表示“time”是“时间”,语义单元表示“the time”是“当代”,语义单元表示“The Times”是“泰晤士报”.

又如,句子 2:“The committee consists of scientists and engineers”是由 5 个基本语义单元表示

“ \langle noun \rangle consist of \langle noun \rangle ”、“ \langle noun \rangle and \langle noun \rangle ”、“the committee”、“scientist”和“engineer”通过代入形成的.

即句子 2 在英语上的句子表达式 S_E-exp 是“‘the committee’ consist of ‘scientist’ and ‘engineer’”.

再如,句子 3:“How do you do”是基本语义单元表示,其对应的语义单元是基本语义单元.

3.4 三种语义单元及语义单元表示

3.4.1 不带类型变量语义单元表示 $S-R_I SU$

不带类型变量的语义单元表示 $S-R_I SU(I=1, 2, \dots)$ 就是没有类型变量的语义单元表示. 显然, I -自然语言上不带类型变量的语义单元表示 $S-R_I SU$ 就是 I -自然语言终结符字母表上的字,即 $S-R_I SU \in (T_I)^*$. 也就是, I -自然语言上不带类型变量的语义单元表示 $S-R_I SU$ 的全集 $S_I \subset (T_I)^*$. 其中 S 表示简单. $S_I \neq (T_I)^*$ 是显然的.

在第 4、5 节中,所有不带类型变量的叶节点都

是不带类型变量的语义单元表示.

3.4.2 一般带类型变量的语义单元表示 $G-R_I SU$

一般语义单元表示就是带有类型变量(由非终结符来表示)、同时也有 I -自然语言的终结符的语义单元表示. 显然, I -自然语言上一般语义单元表示 $G-R_I SU (I=0, 1, 2, \dots)$ 就是 I -自然语言终结符字母表加类型表上的字, 即 $G-R_I SU \in (T_I \cup N)^*$. 也就是, I -自然语言上一般语义单元表示 $G-R_I SU$ 的全集 $G_I \subset (T_I \cup N)^*$. 其中 G 表示一般. $G_I \neq (T_I \cup N)^*$ 也是显然的.

在第 4、5 节中, 所有带类型变量的叶节点都是带类型变量的语义单元表示.

3.4.3 抽象语义单元表示 $A-R_I SU$

某些 I -自然语言上的某些抽象语义单元表示 $A-R_I SU (I=0, 1, 2, \dots)$ 只是有类型变量. 即, 只有非终结符表(即类型表)上的字, 即 $A-R_I SU \in (N)^*$, 没有任何终结符, 其中 A 表示抽象.

例如, “ $\langle \text{num.} \rangle \langle \text{noun} \rangle$ ” 或者 “ $\langle \text{num.} \rangle \langle L_{\text{形式量词}} \rangle \langle \text{noun} \rangle$ ”; 又如, “ $\langle \text{adj} \rangle \langle \text{noun} \rangle$ ”, “ $\langle \text{noun} \rangle \langle \text{noun} \rangle$ ”. 其中, $\langle x \rangle$ 表示非终结符, $L(x)$ 表示 x 的函数. 具体例子如: ‘three’ ‘sheeps’, ‘3’ ‘头’ ‘牛’, ‘兵器’ ‘博物馆’.

有些 I -自然语言上的抽象语义单元表示 $A-R_I SU$ 是 I -自然语言终结符字母表加非终结符表上的字, 即 $A-R_I SU \in (T_I \cup N)^*$. 也就是, I -自然语言上抽象语义单元表示 $A-R_I SU$ 的全集 $A_I \subset (T_I \cup N)^*$.

例如, “ $\langle \text{noun}_1 \rangle \text{ of } \langle \text{noun}_2 \rangle$ ”, “ $\langle \text{noun}_2 \rangle$ 的 $\langle \text{noun}_1 \rangle$ ”.

注意: 因为类型变量在各自然语言中次序不全相同, 甚至还有交错现象, 所以即使 $A_I \subset (N)^*$, 对不同自然语言, 即 $I \neq J$ 时, 仍然可能有 $A-R_I SU \neq A-R_J SU$.

在第 4~6 节中, 部分带类型变量的叶节点是抽象语义单元表示.

3.5 基本语义单元及基本语义单元表示

上面已经提到, 不能由其它语义单元(或者语义单元表示)代入产生的语义单元(或者语义单元表示)被称作基本语义单元(或者语义单元表示). 基本语义单元(或者语义单元表示)同样有 3 种:

不带类型变量的基本语义单元及其表示 $S^0-R_I SU$ 的全集 $S_I^0 \subset S_I \subset (T_I)^*$, $I=0, 1, 2, \dots, n$.

一般带类型变量的基本语义单元及其表示 $G^0-R_I SU$ 的全集 $G_I^0 \subset G_I \subset (T_I \cup N)^*$, $I=0, 1, 2, \dots, n$.

抽象基本语义单元及其表示 $A^0-R_I SU$ 的全集

$A_I^0 \subset A_I \subset (N)^*$, 或者 $A_I^0 \subset A_I \subset (T_I \cup N)^*$, $I=0, 1, 2, \dots, n$.

其中 $R_0 SU$ 为语义单元 SU .

以下讨论的语义单元及其表示都是指基本语义单元及基本语义单元表示. 以下的类型变量都可以是扩充类型变量.

3.6 歧 义

我们称两个或者两个以上的不同语义单元在 I -自然语言上的同一表示为歧义.

当两个或者两个以上的不同语义单元在一种自然语言上有一个相同的表示时, 即存在歧义. 对其中部分可以通过一些附加信息(例如领域信息、场所信息)来区分. 我们可以把这些附加信息作为语义单元表示的一部分来消除歧义. 在人工或者机器翻译过程中出现不可消除的歧义时, 就不能得到唯一确定的正确翻译, 只能得到几种参考译文.

3.7 不同自然语言的语义单元表示的缺位和近似对应

部分语义单元对应的表示不是各自然语言上都有. 没有其语义单元表示的, 称为缺位. 例如, 叶公好龙这个语义单元, 其表示在非汉语的类型结构树上缺位的. 缺位可以填补, 也可以不补.

相当多的语义单元在其它自然语言上对应的表示不都是一致的, 只是相近. 对相近的严格处理可以使用缺位来表示, 非严格处理可以采用近似. 例如, 汉语的“笔”和英语的“pen”相近, 对应的树结构如下:

笔/ (粉笔/chalk,
/pen (钢笔/fountain pen,
圆珠笔/ball pen,
铅笔/pencil,
蘸水笔/dip pen,
...)
)

4 类型树(Category-tree)

类型树是由叶节点和类型节点组成, 树中的非叶节点为类型节点.

类型树不是传统意义下的树, 是准树. 它满足以下特点:

(1) 类型树中的一个类型节点可以有一个或者多个下位节点;

(2) 类型树中的一个叶节点或者非叶节点都可以有多个上位节点;

(3) 一个节点的上节点或者多层的上节点不可

以是其下节点或者多层的下节点。

其中,该节点的 2-层子节点是指一个节点的下位节点;该节点的多层下位节点是指该节点的 2-层或者两层以上的下位节点;该节点的 2-层上位节点是指一个节点的上位节点;该节点的多层上位节点是指该节点的 2-层或者两层以上的上位节点。

(4) 当一个类型节点的子节点出现有叶节点时,该类型节点的其它子节点仍然可以是类型节点。

在本文中,类型节点加上尖括号“ \langle ”,“ \rangle ”,以示区别。

一个类型节点的全部下位节点及全部多层下位节点被称作为下位子树。一个类型节点的全部 $j \leq k$ 的 j 的 j -层下位节点被称为下位 k -层子树,或者下位部分子树。一棵类型树的初始类型节点开始的下位 k 层子树又被称作上 k -层子树。

4.1 事物树

事物树的上若干层子树的节点一般是以类型节点为主。

例如,上 3 层子树:

\langle 事物 \rangle (事物,
 \langle 事件 \rangle (事件, \langle 自然事件 \rangle (自然事件, \langle 自然无生命事件 \rangle (...), \langle 自然有生命事件 \rangle (...)),
 \langle 人造事件 \rangle (人造事件,
 \langle 人造无生命事件 \rangle (...),
 \langle 人造有生命事件 \rangle (...))),
 \langle 物 \rangle (物, \langle 人造物 \rangle (人造物, \langle 有生命人造物 \rangle (...),
 \langle 无生命人造物 \rangle (...)),
 \langle 自然物 \rangle (自然物, \langle 有生命自然物 \rangle (...),
 \langle 无生命自然物 \rangle (...)))
 \rangle).

又如,

\langle 无生命自然物 \rangle (无生命自然物, \langle 无生命自然实体 \rangle (...),
 \langle 抽象物 \rangle (抽象物,
 \langle 科学 \rangle (科学,
 \langle 自然科学 \rangle (自然科学,
 \langle 数学 \rangle (...), \langle 物理 \rangle (...),
 \langle 化学 \rangle (...), \langle 天文 \rangle (...),
 \langle 地理 \rangle (...), \langle 生物 \rangle (...),
 \langle 气象 \rangle (...),...),
 \langle 社会科学 \rangle (...)),
 \langle 自然概念 \rangle (...), \langle 属性 \rangle (...),
 \langle 思想 \rangle (...), \langle 数量 \rangle (...),
 \langle 自然关系 \rangle (...), \langle 空间 \rangle (...), \langle 时间 \rangle (...),
 \langle 度量 \rangle (\langle 定量度量 \rangle (...),
 \langle 定性度量 \rangle (...)),
...)

).

事物树的节点无论是类型节点或者是叶节点,都有带类型变量和不带类型变量两种。不带类型变量的类型节点的下位节点可以是带类型变量的类型节点,也可以是不带类型变量的类型节点,还可以同时有带类型变量的类型节点和不带类型变量的类型节点。

例如,

\langle 姓名 \rangle (\langle 名词_{name with title} \rangle (\langle 姓名 \rangle (\langle 名词_{jobs and professions title} \rangle
 \langle 姓名 \rangle 教授,(...),...),
 \langle 姓 \rangle (\langle 名 \rangle (张军,李白,王小二,(...),
 \langle 名 \rangle (\langle 男性名 \rangle (晓军,(...), \langle 女性名 \rangle (莉莉,(...)),
 \langle 姓 \rangle (白,(...),
...)).

带类型变量节点的下位节点都带有与上位节点相同类型变量,或者相同类型的子类型变量。

例如,

上位节点: \langle 姓名 \rangle (\langle 家庭关系称呼名词 \rangle);
下位节点: \langle 女性姓名 \rangle 姑妈,姑父(\langle 男性姓名 \rangle ,
 \langle 女性姓名 \rangle 舅妈, \langle 男性姓名 \rangle 舅舅,(...).

其中, \langle 姓名 \rangle 、 \langle 家庭关系称呼名词 \rangle 、 \langle 女性姓名 \rangle 、 \langle 男性姓名 \rangle 等等都是类型变量。

4.2 句 树

句树的上若干层子树的节点是类型节点:

\langle 句 \rangle (\langle 组句 \rangle (\langle [基本句] \rangle_n (基本句)和(基本句),
(基本句) \langle 而 \rangle (基本句),(...),
(基本句)(\langle 存在句 \rangle), \langle 关系句 \rangle), \langle 动作句 \rangle ,
 \langle 变化句 \rangle), \langle 思维活动句 \rangle)
 \rangle).

\langle 动作句 \rangle (\langle 存在变化的动作句 \rangle (...), \langle 性质变化的动作句 \rangle (...),
 \langle 状态变化的动作句 \rangle (...), \langle 关系变化的动作句 \rangle (...),
 \langle 位置变化的动作句 \rangle (...), \langle 所有权变化的动作句 \rangle (...),
 \langle 使用权变化的动作句 \rangle (...), \langle 效益变化的动作句 \rangle (...),
 \langle 认知变化的动作句 \rangle (...), \langle 精神变化的动作句 \rangle (...),
...).

其中, $[X]_n$ 表示 \langle 空 \rangle ,或者 X ,或者 XX ,或者 XXX ,或者...,或者 n 个 X ;动作(action)。

句树的类型节点可以带类型变量,也可以不带类型变量。但是其叶节点一般都是带类型变量的,而且所带的变量数目及变量类型对不同的节点可以不一样,其变量类型可以是“或结构”。

句树的叶节点是动作类的基本语义单元表示。

例如,

“ $\langle N_{\text{people/org./machine}} \rangle$ 把 $\langle N_{\text{thing}} \rangle$ 从 $\langle N_{\text{place}} \rangle$ 搬到 $\langle N_{\text{place}} \rangle$ ”,
“ \langle 名词_{公司|个人} \rangle 控告 \langle 名词_{公司|个人} \rangle ”,其中 \langle 名词_{公司|个人} \rangle 是或结构。
 \langle 位置变化的动作句 \rangle

(\langle 搬的动作句 \rangle ($\langle N_{\text{people/org./machine}} \rangle$ 把 $\langle N_{\text{thing}} \rangle$ 从
 $\langle N_{\text{place}} \rangle$ 搬到 $\langle N_{\text{place}} \rangle$),

$\langle N_{\text{people|org. |machine}} \rangle$ 搬 $\langle N_{\text{thing}} \rangle$ 到 $\langle N_{\text{place}} \rangle$,
 $\langle N_{\text{people|org. |machine}} \rangle$ 搬开 $\langle N_{\text{thing}} \rangle$,
 $\langle N_{\text{people|machine}} \rangle$ 正在搬 $\langle N_{\text{thing}} \rangle$,
 $\langle N_{\text{people|org.}} \rangle$ 搬到 $\langle N_{\text{place}} \rangle$, ...,
 (接收的动作句)($\langle N_{\text{people|org. |machine}} \rangle$ 接收到从
 $\langle N_{\text{place}} \rangle$ 来的 $\langle N_{\text{thing}} \rangle$,
 $\langle N_{\text{people|org. |machine}} \rangle$ 接收到 $\langle N_{\text{thing}} \rangle$, ...,
 ...).

4.3 “或结构”的变量类型

句树或者事物树的节点中所带“或结构”的变量类型不是指向句树或者事物树上的一个类型节点,而是指向句树或者事物树上的两个或者两个以上的类型节点之或。

例如,类型变量 \langle 与媒体有关组织|个人|载体|内容 \rangle 是一个或结构的变量类型.它是由 \langle 与媒体有关组织 \rangle , \langle 与媒体有关个人 \rangle , \langle 与媒体有关载体 \rangle 和 \langle 与媒体有关内容 \rangle 等4个不同类型变量“或”在一起所组成.其中,

\langle 与媒体有关组织 \rangle ,其下位节点有: \langle 广播电台 \rangle 、 \langle 电视台 \rangle 、 \langle 广播站 \rangle 、 \langle 出版社 \rangle 、 \langle 报社 \rangle 等等;

\langle 与媒体有关个人 \rangle ,其下位节点有: \langle 电视台台长 \rangle 、 \langle 电视台节目主持人 \rangle 等等,是属于 \langle 人 \rangle 的分支上的节点;

\langle 与媒体有关载体 \rangle ,其下位节点有: \langle 书 \rangle 、 \langle 报 \rangle 、 \langle 杂志 \rangle 、 \langle 文字光盘 \rangle 等等,是属于 \langle 出版物 \rangle 的分支上的节点;其下位节点 \langle 视听光盘 \rangle 是属于 \langle 视听载体 \rangle 的分支上的节点;

\langle 与媒体有关内容 \rangle 属于 \langle 人工创造物 \rangle ,包括 \langle 新闻 \rangle 、 \langle 评论 \rangle 、 \langle 论文 \rangle 、 \langle 故事 \rangle 、 \langle 小说 \rangle 、 \langle 人物介绍 \rangle 、 \langle 传记 \rangle 、 \langle 绘画 \rangle 、 \langle 音乐 \rangle 、 \langle 塑像 \rangle 等等。

5 限定树的“或-与-或结构”

句子和事物语义单元表示都可以带其对应的限定语义单元表示,限定语义单元表示还可以带其对应的限定语义单元表示。

限定树的上若干层子树的节点是类型节点。

\langle 限定 \rangle (\langle 事物限定 \rangle (\langle 地点 \rangle (\dots) \langle 数 \rangle (\dots) \langle 度量 \rangle (\dots)
 \langle 产地 \rangle (\dots) \langle 品牌 \rangle (\dots)
 \langle 感觉程度 \rangle (\dots) \langle 动作* \rangle (\dots)...
 \langle 颜色 \rangle (\dots) \langle 定性度量 \rangle (\dots)),
 \langle 动作及变化句限定 \rangle (\langle 时间 \rangle (\dots) \langle 频率 \rangle (\dots)
 \langle 程度 \rangle (\dots) \langle 情态 \rangle (\dots)...),
 \langle 关系句限定 \rangle (\langle 时间 \rangle (\dots) \langle 程度 \rangle (\dots) \langle 情态 \rangle (\dots)...),
 \langle 限定的限定 \rangle (\langle 时间 \rangle (\dots) \langle 频率 \rangle (\dots) \langle 程度 \rangle (\dots)
 \langle 情态 \rangle (\dots)...
)).

显然,限定树的第1层的各分支之间是或关系,第2层是与关系,第3层又是或关系.这点与任何分类树(包括事物树和句树在内)不同,分类树的各节点的分支都是或关系。

“与结构”($\langle X \rangle \langle Y \rangle \langle Z \rangle \langle W \rangle$)也可以写成为($\langle X \rangle + \langle Y \rangle + \langle Z \rangle + \langle W \rangle$)。

“或结构”($\langle X \rangle, \langle Y \rangle, \langle Z \rangle, \langle W \rangle$)也可以写成为($\langle X \rangle | \langle Y \rangle | \langle Z \rangle | \langle W \rangle$)。

5.1 事物的限定语义单元表示的“与-或结构”

以下是事物的限定语义单元表示的“与结构”的一个例子:

他看见在桌子上的三盘 古巴 新鲜的 大 芒果.
 他摸到在裤兜里的三个 中国 百年前(生产*)的 小 银毫.
 其中带下横线是限定语义单元表示.在这个例子中,它是由5个限定语义单元表示“在桌子上的”、“三盘”、“古巴”、“新鲜的”、“大”等与在一起所构成.我们可以把这5个语义单元表示作为一个“与结构”的5个分量.它们可以同时出现,或者部分出现,或者都不出现。

\langle (地点) $\rangle + \langle$ (数) $\rangle + \langle$ (度量) $\rangle + \langle$ (产地) $\rangle + \langle$ (品牌) $\rangle + \langle$ (感觉程度) $\rangle +$
 \langle (动作*) $\rangle + \dots + \langle$ (颜色) $\rangle + \langle$ (定性度量) \rangle .

其中,各分量的出现是独立的,可以都出现,也可以部分出现,还可以都不出现。

注: \langle 动作* \rangle 可以是一个作为限定语义单元表示的句子或者语言成分.例如,张三去年建造的,民国时期生产的,委托李四生产的,等等。

以下是事物的限定语义单元的“与-或结构”(“与结构”加上其“或子结构”)的例子:

\langle 事物限定树 \rangle (
 \langle (地点) \rangle (\langle (地点名词) \rangle (公园,橱柜,...) \langle 上下 \rangle
 \langle (上,下,里,外,中,...) $\rangle +$
 \langle (数) \rangle (1,2,1.5,...) $+$
 \langle (度量) \rangle (\langle (精确度量) \rangle (公斤,米,...),
 \langle (不精确度量) \rangle (碗,瓶,盘,...),
 \langle (形式度量) \rangle (匹,头,...) $\rangle +$
 \langle (产地) \rangle (亚洲,非洲,中国,美国,北京,天津,...) $+$
 \langle (品牌) \rangle (海尔,大华,...) $+$
 \langle (感觉) \rangle (使人愉快的,令人感兴趣的,令人惊讶的,...) $+$
 \langle (动作) \rangle (\dots) $+$
 $\dots +$
 \langle (颜色) \rangle (红,黄,绿,...) $+$
 \langle (定性度量) \rangle (\langle (新旧) \rangle (新,旧,...), \langle (粗细) \rangle (粗,细,...),
 \langle (大小) \rangle (超大,大,中,...),
 \dots).

其中,“(公园,橱柜,...)”,“(上,下,里,外,中,...)”,“(精确度量), \langle 不精确度量 \rangle , \langle 形式度量 \rangle ”,

“(公斤,米,…)”,等等都是“或结构”。

5.2 动作句、变化句等的限定语义单元表示的“与结构”

动作句、变化句等的限定语义单元表示的“与结构”如

(〈时间〉+〈频率〉+〈动作定性度量〉+〈情态〉+…).

动作限定语义单元的“与向量”的各分类可以独立出现,也可以都出现或者都不出现。

5.3 动作句、变化句等的限定语义单元表示的“与-或结构”

动作句、变化句等的限定语义单元的“与结构”中加上“或子结构”组成“与-或结构”,如

(〈时间〉(〈,在〉〈时间点,段〉(春天,五月,昨夜三点正)

〈里外〉(〈,左右,里,以后,以来,…)〉+

〈频率〉(经常,罕见,很少,…) +

〈动作定性度量〉(〈动作程度〉(重重地,轻轻地,

深深地,…) ,

〈动作速度〉(迅速地,慢慢地,…) ,

〈动作性质〉(和平地,野蛮地,…) ,

…) +

〈情态〉(可能,必须,一定,应该,…) +

…) .

其中,“重重地”,“轻轻地”,“深深地”之间是“或子结构”。

以下是事物的限定语义单元的“与-或结构”的一个例子:

他 可能 在昨天夜里 在运动场上 跑步。

他 一定 在清晨 在花园里 慢慢地 跑步。

其中,第1句的〈动作速度〉是空。

5.4 关系句的限定语义单元表示的“与-或结构”

关系句的限定语义单元表示的“与结构”如

(〈时间〉+〈程度〉+〈情态〉+…).

关系的限定语义单元表示的“与结构”的各分类同样可以独立出现,也可以都出现或者都不出现。

关系的限定语义单元表示的“与-或结构”如

(〈时间〉(〈,在〉〈时间点,段〉(春天,五月,昨夜三点正)〈里外)

(〈,左右,里,以后,以来,…)〉+

〈程度〉(至少,至多,…) +

〈情态〉(可能,必须,一定,应该,…) +

…) .

例句:

这条鱼 在昨天 可能 是活的(现在已经死了)。

他们 去年 肯定 至少 是夫妻(可能还是同谋犯)。

5.5 限定的限定语义单元表示的“与-或结构”

限定的限定语义单元表示树的“与-或结构”如

(〈限定的限定〉

(〈限定程度〉(非常,十分,相当,完全,高度,…) +

〈限定〉(好/坏,难/易,正/错,高/低,有趣的,

令人惊叹的,…) +

…))

例如,

他 刚刚 完成 一道 十分难 相当有趣 非常令人惊叹 的 数学题。

他 将 攀登 一座 十分寒冷 相当陡峭 非常高 的 冰山。

6 扩充类型变量与“选择与次序置换函数”

在以下产生式讨论中,我们将用扩充类型变量 $F_I(B_0, A)$ 代替原来的语义单元表示所带的类型变量 A ,使得各种类型变量可以带上各种合法的限定定义及其表示(即限定词)。扩充类型 $F_I(B_0, A)$ 是原来类型变量 A 和它的限定类型变量 B_0 的函数,该函数是“或-与-或结构”的限定类型 B_0 子集的“选择和次序的置换”函数,该子集的选择是由被限定的类型变量 A 所确定,次序的置换是由自然语言 I 所确定。

6.1 事物树上的节点的扩充类型变量

事物树上节点的扩充事物类型变量为 $F_{I_{\text{事物}}}(B_0, A) = BAB_1$,其中, A 为事物类型变量; B_0 是限定类型变量;限定类型变量 B_0 还可以带其限定类型变量; B 和 B_1 是 B_0 及其所带的限定类型变量的子集。

6.2 句树上的节点的扩充类型变量

句树上的节点的扩充类型变量为 $F_{I_{\text{句}}}(B_0, A)$,其中, A 为动作类变量型; B_0 是动作的限定类型变量。

在动作句和变化句的情况下,动词与其限定词,作为字符串,其次序(与自然语言 I 有关)可能是交错的。

在关系句的情况下, $F_{I_{\text{句}}}(B_0, A) = BA$, B 是 B_0 及其所带的限定类型变量的子集,也就是限定类型变量 B 还可以带其限定类型变量。

6.3 限定树上节点的扩充类型变量

扩充限定类型变量为 $F_{I_{\text{限定}}}(B_0, A)$,其中, A 为限定类型变量; B_0 是限定类型变量的限定类型变量。

6.4 汉语的形式量词和英语的群量词与普通函数

例如,汉语的“一 L (马) 马”,“一 L (牛) 牛”,其中,函数 $L(\text{马}) = \text{匹}$, $L(\text{牛}) = \text{头}$,都是没有实际语义的形式量词。

英语的“a Q(fish) of fish”, “a Q(building) of buildings”, 其中, 函数 $Q(\text{fish}) = \text{school}$, $Q(\text{building}) = \text{cluster}$ 都是表示一群的群量词.

上述的单纯的普通函数关系可以用上下文有关文法来描述, 但是与分量取空, 选择与次序置换函数 $F_I(B_0, A)$ 不能用上下文有关文法进行描述, 是属于 0-型文法.

7 I-自然语言的生成式集 P_I

生成 I-自然语言句子的生成式集 $P_I, I=0, 1, 2, \dots, n$ 有以下 6 种:

(1) 从初始符到句树的根的生成式.

(2) 所有的从句树全部上位类型节点(包括带或者不带类型参量的类型节点(作为左部))到其下位节点, 直到其叶节点的部分.

(3) 所有的从事物树全部上位类型节点(包括带或者不带类型参量的类型节点(作为左部))到其下位节点, 直到其叶节点的部分.

(4) 限定语义单元表示中的“与结构”的各个类型节点到“空”的生成式.

(5) 限定语义单元表示中的“或结构”的到“或结构”的一个分量的生成式.

(6) 从选择与次序置换函数到其取值的生成式.

7.1 一个从初始符到句树的根的生成式

一个从初始符到句树的根的生成式:

$$S \rightarrow \langle \text{句树的根} \rangle.$$

显然, 这个生成式的右部是对各自然语言都相同的非终结符. 实际上, 上几层都是“类型节点”, 各自然语言都是一样.

7.2 句树和事物树的上位类型节点到下位节点的生成式集

生成式集包括两部分.

所有的从句树全部上位类型节点(包括带或者不带类型参量的类型节点(作为左部))到其下位节点, 直到其叶节点部分. 其叶节点一般都是带类型变量, 所带的类型变量数目及类型变量的类型可以各自不一样, 而且其类型变量可以是“或结构”的类型变量, 句树的叶节点是动作类的带类型变量的基本语义单元表示(参见 4.2 节句树).

所有的从事物树全部上位类型节点(包括带或者不带类型参量的类型节点(作为左部))到其下位节点, 直到其叶节点部分. 其叶节点可以带类型变量(一般是不带的), 所带的类型变量的类型可以各自

不一样, 而且可以是“或结构”, 事物树的叶节点是带类型变量或者不带类型变量的基本语义单元表示(参见 4.1 节事物树).

$\langle \text{事物} \rangle \rightarrow \text{事物}; \langle \text{事物} \rangle \rightarrow \langle \text{事件} \rangle; \langle \text{事物} \rangle \rightarrow \langle \text{物} \rangle; \dots$

$\langle \text{物} \rangle \rightarrow \text{物}; \langle \text{物} \rangle \rightarrow \langle \text{自然物} \rangle; \langle \text{物} \rangle \rightarrow \langle \text{人造物} \rangle; \dots$

$\langle \text{人造物} \rangle \rightarrow \text{人造物}; \langle \text{人造物} \rangle \rightarrow \langle \text{有生命的人造物} \rangle;$

$\langle \text{人造物} \rangle \rightarrow \langle \text{无生命的人造物} \rangle;$

$\langle \text{名词}_{\text{name with title}} \rangle \rightarrow \langle \text{名词}_{\text{family relationship title}} \rangle \langle \text{姓名} \rangle;$

$\langle \text{名词}_{\text{name with title}} \rangle \rightarrow \langle \text{名词}_{\text{high rank title in society}} \rangle \langle \text{姓名} \rangle; \dots$

$\langle \text{名词}_{\text{family relationship title}} \rangle \langle \text{姓名} \rangle \rightarrow \text{舅妈} \langle \text{女性姓名} \rangle;$

$\langle \text{名词}_{\text{family relationship title}} \rangle \langle \text{姓名} \rangle \rightarrow \text{舅舅} \langle \text{男性姓名} \rangle; \dots$

$\langle \text{姓名} \rangle \rightarrow \langle \text{姓} \rangle; \langle \text{姓名} \rangle \rightarrow \langle \text{名} \rangle; \langle \text{姓名} \rangle \rightarrow \langle \text{姓} \rangle \langle \text{名} \rangle; \dots$

$\langle \text{名} \rangle \rightarrow \langle \text{女性的名} \rangle; \langle \text{名} \rangle \rightarrow \langle \text{男性的名} \rangle; \dots$

$\langle \text{男性的名} \rangle \rightarrow \text{军}; \langle \text{女性的名} \rangle \rightarrow \text{莉莉}; \dots$

7.3 限定语义单元表示中节点到空或者到一个分量的生成式集

生成式集包括两部分.

限定语义单元表示中的“与结构”的各个类型节点到“空”的生成式; 限定语义单元表示中的“或结构”的到“或结构”的一个分量的生成式.

例如, 或类型 $A = B | C | D; A \rightarrow B, A \rightarrow C, A \rightarrow D$ 3 个生成式是需要的.

7.4 类型变量中的“或类型”到其所包含的各自的类型参量的生成式集

生成式集包括两部分.

从选择与次序置换函数到其取值的生成式集; 包括选择与次序置换的生成式集.

7.5 基本生成式

基本生成式是指不可以用两个或者两个以上的生成式所产生的生成式.

例如, “ $\langle \text{名称谓的人} \rangle \rightarrow \text{史密斯教授}$ ”不是基本生成式, 因为它可以由生成式:

$\langle \text{名称谓的人} \rangle \rightarrow \langle \text{名称谓} \rangle \langle \text{姓名} \rangle$, “ $\langle \text{名称谓} \rangle \langle \text{姓名} \rangle \rightarrow \langle \text{名称谓} \rangle \langle \text{姓名} \rangle$ ”,

$\langle \text{名称谓} \rangle \langle \text{姓名} \rangle \rightarrow \langle \text{姓名} \rangle \text{教授}$, “ $\langle \text{姓名} \rangle \rightarrow \langle \text{姓} \rangle$ ”, 和 “ $\langle \text{姓} \rangle \rightarrow \text{史密斯}$ ”

产生.

无论是事物树或者句树, 一个类型节点到其下下位节点都不是基本生成式.

在生成式集 P_I 中, 只包含基本的生成式. 因为非基本的生成式可以由基本生成式代入所产生.

注: 本文书写过程中参考的是英语词汇及语法书^[12-17].

7.6 例子

$S \rightarrow \langle \text{句树的根} \rangle \rightarrow \langle \text{动作句} \rangle \rightarrow \langle \text{所有权变化的动作句} \rangle$

→〈人〉卖给〈人〉〈物〉
 →〈姓名〉卖给〈带头衔姓名〉〈带限定语的物〉
 →〈姓名〉卖给〈姓名〉〈头衔〉〈限定语〉〈物〉
 →〈姓名〉卖给〈姓名〉〈头衔〉〈精确度量〉〈产地〉〈品牌〉
 〈感觉〉〈颜色〉〈定性度量〉〈物〉
 →〈姓名〉卖给〈姓名〉〈头衔〉〈精确度量〉〈产地〉〈品牌〉
 〈感觉〉〈颜色〉〈定性度量〉〈物〉
 →张三卖给李博士十公斤烟台大苹果。
 该例子使用了以下的 22 个生成式:
 S→〈句树的根〉;
 〈句树的根〉→〈动作句〉;
 〈动作句〉→〈所有权变化的动作句〉;
 〈所有权变化的动作句〉→〈人〉卖给〈人〉〈物〉;
 〈人〉→〈带头衔姓名〉;
 〈带头衔姓名〉→〈姓名〉〈头衔〉;
 〈物〉→〈精确度量〉〈产地〉〈品牌〉〈感觉〉〈颜色〉
 〈定性度量〉〈物〉;
 〈姓名〉→张三;
 〈头衔〉→博士;
 〈姓名〉→〈姓〉;
 〈姓〉→〈李〉;
 〈精确度量〉→〈数〉〈精确度量名词〉;
 〈数〉→十;
 〈精确度量名词〉→公斤;
 〈产地〉→烟台;
 〈品牌〉→〈空〉;
 〈感觉〉→〈空〉;
 〈颜色〉→〈空〉;
 〈定性度量〉→〈大小〉;
 〈大小〉→大;
 〈物〉→〈果类食物〉;
 〈果类食物〉→苹果。

8 结 论

基于自然语言语法规则由 Chomsky 的上下文无关文法所生成的句子合乎语法规则,但是不一定合乎语义常理. 基于语义单元及其表示理论由 Chomsky 的短语文法构成的生成系统 $LRSGS_I$ 所生成的 I -自然语言上的句子是既合法又合理. 其特点是各自然语言统一,基于语言单元理论且具有数以百万计生成规则,即比较精确描述自然语言的文法是短语文法,其生成句子“合理”的关键是基于语义单元理论并引入了“类型参量”系统. 类型参量系统的作用不仅体现在语义单元集 B_I 中,而且还体现在类型树 C_I 上和函数集合 F_I 中.

参 考 文 献

- [1] Chomsky N. Three models for the description of language. IRE Transactions on Information Theory, 1956, 2: 113-124
- [2] Katz J J. Semantic Theory. New York: Harper and Row, 1972
- [3] Fillmore C J. The case for case reopened//Cole P, Sadock J M ed. Syntax and Semantics 8, Grammatical Relation. New York: Academic Press, 1977
- [4] Tesnière Lucien. Eléments de syntaxe structurale (Foundation of Structural Syntax). Ed. by Fourquet J and Daumas F. Paris: Klincksieck, 1959
- [5] Gao Qing-Shi, Chen Zhao-Xiong. Human-like machine translation. Journal of Computer Research and Development, 1989, 26(2): 1-8(in Chinese)
(高庆狮, 陈肇雄. 类人机译原理. 计算机研究与发展, 1989, 26(2): 1-8)
- [6] Gao Q S, Gao X Y, Hu Y. Semantic language and multi-language MT approach based on SL. Journal of Computer Science & Technology, 2003, 18(6): 848-852
- [7] Gao Xiao-Yu, Gao Qing-Shi, Hu Yue. High speed multi-language machine translation approach based on pruning on the tree representations of semantic elements. Journal of Software, 2005, 16(11): 1000-1919(in Chinese)
(高小宇, 高庆狮, 胡玥. 基于语义单元表示树剪枝的高速多语言机器翻译方法. 软件学报, 2005, 16(11): 1000-1919)
- [8] Gao X Y, Hu Y, Li L, Gao Q S. Semantic-parsing based on semantic units theory—A new approach to natural languages processing. International Journal of Pattern Recognition and Artificial Intelligence, 2008, 22(11): 1447-1459
- [9] Hu Yue, Gao Qing-Shi, Gao Xiao-Yu. United algorithm base of multi-language machine translation system: Difference function. Journal of Beijing University of Aeronautics and Astronautics, 2008, 34(5): 604-608(in Chinese)
(胡玥, 高庆狮, 高小宇. 多语互译机译系统统一算法基础——差异函数. 北京航空航天大学学报, 2008, 34(5): 604-608)
- [10] Hu Yue, Gao Xiao-Yu, Li Li, Gao Qing-Shi. Style transformation in natural languages. Frontiers of Computer Science in China, 2008, 2(4): 424-430
- [11] Gao X Y, Hu Y, Li L, Gao Q S. Semantic-parsing based on semantic units theory. International Journal of Pattern Recognition and Artificial Intelligence, 2008, 22(7): 1447-1459
- [12] Fellbaum C. Word-Net: An Electronic Lexical Database. Boston: MIT Press, 1998
- [13] Stephen G. Word Menu. New York: Random House, 1998
- [14] Barbara A K. Rogets. 6th Edition. New York: Harper Collins Publishers Ltd, 2001
- [15] Sinclair J ed. In chief. Collins COBUILD English Language Grammar. New York: Harper Collins Publishers Ltd, 1990 (English-Chinese, Hong Kong: The Commercial Press, 1999)

- [16] Sinclair J ed. In chief. Collins COBUILD Grammar Patterns 1: Verbs. New York: Harper Collins Publishers Ltd, 1992
- [17] Sinclair J ed. In chief. Collins COBUILD Grammar Patterns

2: Nouns and Adjectives. New York: Harper Collins Publishers Ltd, 1992



HU Yue, born in 1963, Ph. D., associate professor. Her research interests include nature language processing, computer architecture, network security.

GAO Xiao-Yu, born in 1978, M. S.. His research interests include natural language processing, network security

Background

In the natural language processing, it is important that how to estimate whether a sentence is legitimate and reasonable. Current estimation tool of natural language is based on syntax and by Chomsky's Context-free Grammar, it can't estimate whether the sentence is reasonable.

The sentences produced by new generating system, proposed in this paper, are legitimate and reasonable. i. e., it can estimate that whether the sentence is reasonable and legitimate. The feature of new generating system is based on semantic unit theory and the number of rule is several millions, rather than several thousands, and by Phrase-structure Grammar rather than by Context-free Grammar, i. e. the grammar to describe more exactly natural language must be Phrase-structure Grammar, and its productions are based on semantic units theory. The key of reasonableness is based on semantic-units theory, to introduce the category-parameter system, which is in the set $B-R_1SU$, in the set $C-tree_1$, and in the set F_1 .

This work is belong to a subject of natural language pro-

and string match.

LI Li, born in 1980, Ph. D., lecturer. Her research interests include natural language processing and string match.

GAO Qing-Shi, born in 1934, academician of Chinese Academy of Sciences. His research interests include parallel algorithm, computer architecture, natural language and machine translation, human intelligence, network security and string match.

cessing, and was supported by Grant No.2007CB311103 from the 973 Program of China, Grant Nos.2006AA01Z140, 2006AA010101 from the 863 Program of China, and the Grant Nos.60873002, 60736013 from the National Science Foundation of China (NSFC).

The team had establish new theory of unified natural linguistics, include analyzing natural linguistics as a natural science, unified natural linguistics by transformation, style transformation between different components inner one language, and corresponding transformation between different natural language, unified processing approach in different natural language, etc.

The team had been publish the book, "Foundations of Unified Linguistics", in Science Press, and several papers in International journal of pattern recognition and artificial intelligence, Journal of Computer Science & Technology, Journal of Software, Journal of Computer Research and Develop, etc.