汉英机译系统 ICENT 中主语省略句的处理*

马红妹　齐璇　王挺　陈火旺

(国防科技大学计算机学院, 长沙 410073)

E-mail: mhmma@sina.com

摘要: 汉语中主语省略现象十分普遍,汉语主语省略句的处理对于汉英机器翻译十分重要,它需要基于篇章上下文语境进行分析,包括省略主语识别和省略主语恢复.本文首先介绍了汉英机译系统 ICENT 的句法语义分析,然后建立了汉语篇章上下文语境模型,制定了主语省略恢复规则,给出了基于汉语篇章上下文语境应用主语省略恢复规则恢复主语省略的算法,最后对小学语文课本实际语料进行了实验.

关键词: 汉英机器翻译; 语义分析; 上下文语境; 主语省略; 主逻辑主语

引言

汉语主语省略句的处理是汉英机器翻译研究的一个难点问题.省略主语的恢复对汉英机器翻译十分重要.它的处理超出了单句范围,需要基于篇章上下文语境进行分析.近几年,国外学者对篇章上下文知识的表示进行了研究[1,2],并讨论了省略、指代等问题;国内学者也对汉语的省略现象进行了有益的探索[3,4].我们面向汉英机器翻译,基于汉语篇章上下文语境研究了汉语主语省略句的处理.

基于中间语言的机器翻译模式具有结构清晰、易于扩充的优点,但是中间语言的设计比较复杂.我们针对汉语和英语两种语言的语言现象设计了一种中间语言,并实现了一个基于中间语言的汉英机译系统 ICENT,它对汉语句子的分析包括句法分析和语义分析[5,6].以此为基础,我们在 ICENT 系统中对汉语主语省略句进行了处理(如图 1 所示).

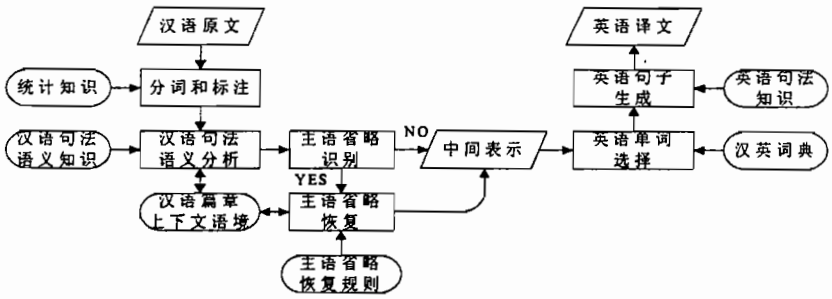

图 1　ICENT 系统中汉语主语省略句的处理过程

处理过程包括主语省略识别和主语省略恢复,主语省略识别判断句子是否省略了主语,主语省略恢复

* 本项目受到国家八六三计划资助(2001AA114110).

根据汉语篇章上下文语境和主语省略恢复规则恢复句子省略的主语.在汉语句子的语义结构中,主谓词是核心,主谓词的逻辑主语又称主逻辑主语表示事件或状态的主体,可直接转换为英语译文的主语.主逻辑主语的省略是对汉英机器翻译影响较大的一种主语省略,因此我们主要讨论了汉语句子语义结构中主逻辑主语省略的识别和恢复.

汉语主语省略多发生在小句中[3],因此我们以小句作为研究对象,分析小句中主语省略问题.我们把小句界定为由标点符号分隔的一串字符串,其语法功能等同于单句或分句.按照省略主语参照点在上下文中的位置,小句主语省略分为承前省略和蒙后省略,本文主要讨论小句主语承前省略的情况.

本文组织如下.第一节介绍了汉语语义分析;第二节建立汉语篇章上下文语境模型;第三节给出了中汉语主语省略句处理方法并在 ICENT 系统中进行了实验;第四节进行了总结.

1 汉语句法语义分析

与英语相比,汉语是一种语义型语言,仅从句法结构上分析汉语会产生大量的歧义结构,因此汉语分析尤其需要语义知识,需要通过语义分析消除歧义.基于对客观世界的认识,我们建立了汉语语义知识表述体系[6],用语义单元、语义角色关系和语义规则来描述汉语语义知识.语义单元是语义知识的载体.语义角色关系描述了概念间、复合概念间的语义关系,我们定义了 56 种语义角色关系.根据语义角色关系的不同,将概念或复合概念分为从属谓词、逻辑主语、逻辑宾语、修饰语和中心语五类,担当不同的语义角色.语义规则是语义知识的抽象,确立语义单元间的合理组合,并描述语义单元间的语义角色关系.句法和语义是形式和内容的关系,二者密不可分,句法分析和语法分析也是相互结合的.汉语的分析采用扩展的上下文无关文法,每一条句法产生式都对应一个前提判断函数.当分析器要用句法产生式进行归约时,首先激活前提判断函数,在其中调用相应的语义分析模块进行语义分析,只有通过语义分析才可以进行归约,否则当前分析不正确,可以终止.在进行规约时,不但产生了新的句法结构,还生成了与之对应的语义结构.分析结束时得到句子的语义结构和合乎语义的句法结构.句子的语义结构由各级语义单元及其语义角色关系构成,用嵌套的框架表示,描述了句子语义的层次性和结构性.

2 汉语篇章上下文语境模型

汉语篇章中,上下文语境指由篇章的词、句构成的内部环境.一方面词语或句子的语义依赖于上下文语境来确定;另一方面,上下文语境又是词语或句子的语义动态积累的结果[7].上下文语境不仅包括句子的语义,还包括句子之间的语义关联.

由于句子的语义可以由其内部概念信息体以及概念信息体之间的语义角色关系表示,句子之间的语义关联也可以由跨语句概念信息体之间的语境连贯关系来等价地表示,因此由句子语义动态构造的上下文语境也可以由概念信息体和概念信息体之间的关系来表示.于是我们建立了汉语篇章上下文语境模型,它由汉语篇章中的概念信息体和概念信息体之间的语义角色关系和语境连贯关系构成.我们采用 CIURN(Concept Information Unit Relation Network, CIURN)来表示汉语篇章的上下文语境,CIURN 的定义如下:

定义 1. 概念信息体关联网络 CIURN

CIURN ::= < CIU, SemRoleRela, ConCohrRela >.

① CIU 是结点集,其元素称为概念信息体,分为基本概念信息体和复合概念信息体.基本概念信息体表示简单概念;复合概念信息体表示概念复合体,其语义由成分概念的语义复合而成.

② SemRoleRela 是有向边集,表示概念信息体之间的语义角色关系集.

③ ConCohrRela 是有向边集,表示概念信息体之间的语境连贯关系集.

CIURN 是由结点和连接结点的弧构成的有向图. 结点表示概念信息体,由复杂特征集表示;弧是有向边,表示概念之间的语义关系,包括语义角色关系和语境连贯关系. 传统的语义网络缺乏表示语义组合性的机制,我们在 CIURN 中引入了复合概念信息体结点表示概念信息体的组合,使得语义网络具有了结构性,这样CIURN也可以表示句子的语义,由句子语义动态建立篇章上下文语境的过程便表示为CIURN网络动态增长的过程.

3 主语省略的识别和恢复

3.1 主语省略识别

主语省略识别判断小句是否省略了主语. 在汉英机器翻译中,汉语小句语义结构中的主谓词和主逻辑主语可用于生成英语译文的谓语和主语. 主逻辑主语的省略是对英语译文主语生成影响较大的一种主语省略,为此本文的主语省略识别主要是识别小句的语义结构中是否省略了主逻辑主语.

3.2 主语省略恢复规则

3.2.1 基本原理

为简便起见,我们称主语省略句中省略的主逻辑主语为省略主语,称省略主语回指的上文对象为省略主语所指. 汉语主语省略现象具有连续性规律,分为微观连续性和宏观连续性[8]. 微观连续性规律说明省略主语所指应该是上文中启后性较强的名词性成分,在小句的语义结构中主要指实体类概念信息体,包括主逻辑主语、实体类逻辑宾语、主逻辑主语的所属修饰语、处所修饰语中的实体类概念和时间修饰语中的逻辑主语类概念,我们称这些概念信息体为省略主语参照点(referent). 宏观连续性规律说明省略主语所指大多位于主语省略句的前后相邻句中,或两个小句在语义结构上间隔一般不超过三个层次.

我们把省略主语所指所在的小句(简称上文小句)到当前主语省略句(简称当前小句)的间隔称为恢复省略主语上下文窗口(简称窗口),窗口中包含小句的个数为窗口的大小. 当窗口大小为 1 时,恢复省略的主语需要判断与当前小句相邻的前一个小句的语义结构;当窗口大小大于 1 时,恢复省略的主语除了需要判断上文小句的语义结构外,还需要判断中间小句的语义信息. 除此之外,恢复省略的主语还与下列因素有关:

(1) 主谓词同形信息

主谓词同形信息包括:上文小句主谓词和当前小句主谓词同形,如"…是…, 是…";上文小句主谓词的修饰语和当前小句主谓词的修饰语同形,如"…一边…, 一边…". 若具有主谓词同形信息,则当前小句承上文小句的主逻辑主语省略.

(2) 上文小句主谓词事件的语义类别

我们把上文小句主谓词按照事件必备角色框架[9]分类. 事件必备角色框架描述了小句应该出现哪些语义角色,若实际并未出现则预示了当前小句可能与上文小句具有语义角色关系,据此可推导当前小句省略的主逻辑主语. 主谓词语义类别主要有目标内容(target-content)类、目标原因(target-cause)类、结果事件(resultevent)类等.

(3) 省略主语参照点语义类别

省略主语参照点均属于实体类概念,语义类别有:人、组织等.

(4) 逻辑主语约束

逻辑主语约束是指事件概念对其逻辑主语语义类别的要求. 省略主语参照点的语义类别满足当前小句主谓词的逻辑主语约束是恢复省略的主逻辑主语的必要条件.

上述信息构成主语省略恢复规则的前提判定条件,主语省略类型为规则的结论.主语省略类型有：承主逻辑主语省略(*LogSubj*)、承实体类逻辑宾语省略(*LogObj*)、承主逻辑主语的所属修饰语省略(*LogSubjPoss*)、承处所修饰语中实体概念省略(*LogModiLEC*)和承时间修饰语中逻辑主语类概念省略(*LogModiTEC*).规则的定义略.

3.2.2 分组与优先级

由于恢复省略主语上下文窗口大小取 2 已覆盖了大多数主语省略现象,因此我们主要就窗口大小为 1、为 2 的情况制定了规则.规则分为两组：A 组和 B 组,A 组是窗口大小为 1 的规则,B 组是窗口大小为 2 的规则.A 组规则不包括中间小句语义信息,其上文小句指当前小句的前一小句.按照同形信息 A 组规则分为两大类:A.1 类规则为同形规则,A.2 类规则为不同形规则.A.2 类规则又按照主语省略类型分为五小类：[A.2.1] *LogSubj*、[A.2.2] *LogObj*、[A.2.3] *LogSubjPoss*、[A.2.4] *LogModiLEC*、[A.2.5] *LogModiTEC*.A 组规则的优先级由高到低依次为：A.1 → A.2.5/A.2.4 → A.2.3 → A.2.2 → A.2.1.B 组规则加入了对中间小句语义信息的判断,其上文小句指当前小句的前一小句的前一小句.B 组规则的具体划分及其优先级同 A 组规则.

3.2.3 实例

由于篇幅有限,本文就 A 组规则和 B 组规则各举一个实例进行说明,[规则 A.2.2.i]是[A.2.2]类规则集中的第 i 条规则,[规则 B.2.5.j]是[B.2.5]类规则集中的第 j 条规则.

[规则 A.2.2.i] ((SemCat (E (结果事件) (行动)) (ER (LS 是) (LO 是))) (SemStru (!SemEle 从属谓词))) → *LogObj*

[规则 A.2.2.i]表明：窗口大小为 1 时,若上文小句主谓词的语义类别是“结果事件”,预示主谓词的实体类逻辑宾语作从属谓词的逻辑主语,当上文小句实际并未出现从属谓词时,则说明上文小句的从属谓词成为当前小句的主谓词,在上文小句主谓词的实体类逻辑宾语满足当前小句主谓词的逻辑主语约束的条件下,当前小句承上文小句主谓词的实体类逻辑宾语省略.如：

例 2：(S_0)警卫员提醒他,(S_1)坐沙发的时候要把腿收回来.

例 2 中,S_0的主谓词“提醒”属“结果事件类”事件概念,主逻辑主语“警卫员”、实体类逻辑宾语“他”均满足 S_1主谓词“说”的逻辑主语约束,S_0中未出现从属谓词,根据[规则 A.2.2.i],得 S_1省略的主逻辑主语是“他”.

[规则 B.2.5.j] ((SemCat (ER (LMT 是))) (IntraClause (Role 时间))) → *LogModiTEC*

[规则 B.2.5.j]表明：窗口大小为 2 时,考虑中间小句语义信息,若中间小句与上文小句或当前小句之间具有“时间”语义角色关系,而且上文小句有且只有时间修饰语,在时间修饰语中的逻辑主语类概念满足当前小句主谓词的逻辑主语约束的条件下,当前小句承上文小句时间修饰语中的逻辑主语类概念省略.如：

例 3：(S_0)列宁 8 岁的时候,(S_1)有一天,(S_2)跟爸爸到姑妈家去做客.

例 3 中,S_0有且只有时间修饰语,其中的逻辑主语类概念“列宁”满足 S_2的主谓词“到”的逻辑主语约束,S_1和 S_2具有“时间”语义角色关系,所以根据[规则 B.2.5.j],得 S_0省略的主逻辑主语是“列宁”.

3.3 主语省略恢复算法

主语省略恢复算法(Zero-anaphoric Subject Recovering Algorithm, ZSRA)算法如下：

算法 1. 主语省略恢复算法 ZSRA

设当前小句为 S_i,当前恢复省略主语上下文窗口大小为 k（阈值为 N）,S_i的上文小句由近及远依次为 S_{i-1}, …, S_{i-k},初始时 k 为 1,S_i的主谓词是 S_i.MP,求 S_i的主逻辑主语 S_i.MP.LogSubj.

Step1 获得 S_i的主谓词 S_i.MP;

Step2 获得 S_i 主谓词 S_i.MP 的逻辑主语约束，设为 S_i.MP.LogSubjDomain；

Step3 从 CIURN 中取出上文小句 S_{i-k}，获得 S_{i-k} 的主谓词 S_{i-k}.MP；

Step4 从 S_{i-k} 获得省略主语参照点 S_{i-k}.MP.LogSubj、S_{i-k}.MP.LogObj、S_{i-k}.MP.LogSubj.Poss、S_{i-k}.MP.LogModiLEC 和 S_{i-k}.MP.LogModiTEC；

Step5 判断省略主语参照点是否满足 S_i.MP.LogSubjDomain；

Step6 如果 k>1，则获取中间小句语义信息；

Step7 应用第 k 组主语省略恢复规则判断主语省略类型，如果有规则命中，设相应的省略主语参照点为 x，则 S_i.MP.LogSubj←x，转 *Step9*；如果没有规则命中，则 k←k+1；

Step8 如果 k ≤ N，则转 *Step3*；否则，S_i.MP.LogSubj←defaultvalue，转 *Step10*；

Step9 在 S_i.MP 和 x 之间建立“省略主语所指”语境连贯关系，由 S_i.MP 指向 x；

Step10 把 S_i 加入篇章上下文语境 CIURN，在 S_i.MP 和 S_{i-1}.MP 之间建立“邻接上下文”语境连贯关系，由 S_{i-1} 指向 S_i；

Step11 算法结束.

ZSRA 应用省略主语恢复规则判断省略主语参照点是否可作为当前小句省略的主语，从而恢复当前小句省略的主逻辑主语，并更新 CIURN，算法 ZSRA 结束. 算法 ZSRA 中设立了恢复省略主语上下文窗口大小的阈值 N，如果窗口大小超过阈值 N 仍没有得到省略的主逻辑主语，则 ZSRA 采用缺省值结束，结束之前将当前小句的 CIURN 加入篇章上下文语境 CIURN，已备下文引用.

我们对小学语文课本前四册中的 43 课文进行了实验，共 1,399 个小句，407 个小句出现主语省略. ZSRA 算法中窗口大小的阈值 N 取为 2，省略主语所指在前一、前二小句的情况基本得到了恢复，约占主语省略句的 85%.

4 结束语

我们从语义的角度对汉语主语省略现象进行了分析和形式化描述，给出了基于篇章上下文语境模型的汉语主语省略句处理方法，并在受限语料中进行了实验，取得了较好的结果. 但是由于受到语义知识资源不足的制约，目前语义分析还有许多地方需要完善；另外汉语中主语省略十分灵活，有些主语省略的恢复不仅需要依据篇章上下文语境，还需要依据更深层次的语境知识，而这些知识的描述至今仍是一个难题，这些问题使得本文的方法还不能面向大规模非受限文本进行测试. 我们下一步工作将继续进行语义知识库的开发，细化语义类别划分标准，扩充主语省略恢复规则集，提高主语省略恢复的准确率.

参考文献：

[1] Eijck J., Kamp H. Representing Discourse in Context. Handbook of Logic and Language. Johan B. and Alice M. ed. MIT press, 1997, 178~237.

[2] 冯志伟. 自然语言机器翻译新论. 北京：语文出版社，1994.

[3] 宋柔. 汉语叙述文中的小句前部省略现象初析. 中文信息学报，1992，6(3)：62~68.

[4] 王厚峰. 汉语省略的判定于恢复研究. HNC 与语言学研究. 武汉：武汉理工大学出版社，2001，236~241.

[5] 周会平. 基于中间语言的汉英翻译系统 ICENT 的研究与实现[博士学位论文]. 国防科学技术大学，1999.

[6] 齐璇. 汉语语义知识的表示及其在汉英机译中的应用[博士学位论文]. 国防科学技术大学，2002.

[7] 马红妹，王挺，陈火旺. 汉英机器翻译中语境知识的表示与应用. 自然语言理解与机器翻译，黄昌宁，张普主编，清华大学出版社，2001，278~284.

[8] 陈平. 汉语零形回指的话语分析. 中国语文，1987(5)，363~378.

[9] 董振东. 知网. http://www.keenage.com

作者简介: 马红妹(1974—), 博士生, 研究方向: 机器翻译、计算语言学; 齐璇(1973—), 博士, 研究方向: 机器翻译、计算语言学; 王挺, 博士, 副教授, 研究方向: 机器翻译、计算语言学和计算机软件; 陈火旺, 中国工程院院士, 教授, 博士生导师.

Processing of Zero Anaphoric Subject of Chinese Sentence in Chinese-English Machine Translation System ICENT*

MA Hongmei QI Xuan WANG Ting CHEN Huowang

(School of Computer, National University of Defence Technology, Changsha 410073, China)

E-mail: mhmma@sina.com

Abstract: Zero-anaphoric subject of Chinese sentence is very prevalent. It is important to process zero-anaphoric subjects of Chinese sentences in Chinese-English machine translation. The process needs to analyze Chinese text based on the knowledge of linguistic context, it includes zero-anaphoric subject recognizing and zero-anaphoric subject recovering. In this paper, syntactic analysis and semantic analysis of Chinese-English machine translation system ICENT are introduced firstly. Then the model of linguistic context of Chinese text is constructed, and zero-anaphoric subject recovering rules are designed. With the linguistic context of Chinese text, Zero-anaphoric subject recovering algorithm (ZSRA) is proposed. Finally, experiment is done using the Chinese textbooks of elementary school.

Key words: Chinese-English machine translation; semantic analysis; linguistic context; zero-anaphoric subject; main logical subject

* Supported by National 863 Foundation No. 2001AA114110