汉语句子规范化处理策略的研究

宗成庆　章森　陈肇雄　黄河燕
(中国科学院计算技术研究所 智能机译中心 北京 100080)

摘要: 汉语句子规范化处理是机器翻译系统中译前处理的重要内容。本文针对口汉语句子的特点和语音识别机制生成句子的特殊性，提出了非规范句子检查和自动校正与规范句型模板约束相结合的人机交互式汉语句子规范化处理策略，并详细讨论了汉语句子分析中的相关问题。

关键字: 句子规范化　译前处理　句型

Research on the Strategies for Standardization Processing of Chinese Sentences

Zong Chengqing　Zhang Sen　Chen Zhaoxiong and Huang Heyan
(IMT Research Center, Institute of Computing Technology, The Chinese Academy of Sciences)

ABSTRACT: In a machine translation system, the standardization processing of Chinese sentences is an important aspect in sentence processing before translation. Considering the specific characteristics of spoken Chinese sentences and the particularity of sentences from the module of Chinese speech recognition, this paper presents strategies for standardization processing of Chinese sentences, which combine the method to find and correct informal sentences automatically by using standard models of sentence patterns with the method to check the results with the help of human-computer interaction. The paper also discusses in detail the problems in Chinese sentence analysis.

KEY WORDS: standardization of sentences, processing before translation, sentence pattern

一、引　言

在机器翻译系统中，句子规范化处理是译前处理的重要内容。非规范句子不仅给机译系统分析机制带来较大困难，而且严重地影响系统的执行效率和译准率。因此，对非规范句子进行译前规范化处理，具有十分重要的意义。

本文论述的是汉英语音翻译系统中非规范句子的译前规范化处理问题。在我们所研究的汉英语音翻译系统中，其分析机制包括语音识别及后处理、句子规范化处理、汉英机器

翻译机制和语音合成四大模块，其基本构成如下图所示：

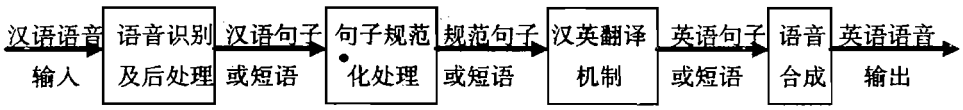

图 1. 汉英语音翻译系统基本构成

由于口汉语中大量地出现非规范句子，而且汉语语音识别及后处理机制也常常由于各种原因而导致非规范句子的产生，因此，句子译前规范化处理在本系统中尤为重要。

关于现代汉语句型，我国众多的专家和学者都做过大量的研究工作。李([1], 1986)和北([2], 1989) 从语言学角度总结了现代汉语的基本句型种类；陈([3], 1986)论述了汉语句型的确定原则和各类句型的基本特点；罗([4], 1994)对现代汉语句型的分布进行了全面的分析和统计；唐([5], 1993)归纳了信息处理用现代汉语简单句的基本类型。所有这些成果都为中文信息处理中的句子处理提供了极有意义的指导和参考。

本文从机器翻译的实用性角度出发，针对口汉语句子的特点和来自语音识别及后处理机制句子的特殊性，提出以非规范句子检查与自动校正和规范句型模板约束相结合，利用人机交互实现汉语句子规范化处理的策略，并详细讨论了句子分析中的相关具体问题。

二、基本定义和约定

由于本文中所用到的几个基本概念与人们日常所指的汉语语法中给出的严格定义略有不同，因此，我们首先对本文中所指的几个基本概念给出粗略的定义或约定。

定义 1：句子 一组连续的汉字字符构成一个句子。

我们通常情况下所说的句子，一般是指能够完整地表达一个意思，具有一定的句法成分的字符串，而在本系统中所指的句子具有广泛的含义，由于它们是汉语口语句子，不象书面语句子之间有标点符号分割，而且在语音识别机制中，句子的划分一般是按相邻音节之间的时间间隔确定的，因此，它可能是一个真正意义上的句子，也可能是一个短语，或一个词，甚至一个字。

定义 2：句型 对具有相同的句法、语义和语用特征的句子的概括和抽象。简单地说，句型就是句子的结构类型。

定义 3：规范句型 根据实际语料客观地统计、归纳、筛选出来的或为特定需要而人为约定的句型。

在本系统中，为了便于句子处理，根据现代汉语口语特点和系统特定的应用领域及其需要，按统计规律总结出来的或人为约定的句型，统称为规范句型。

定义 4：非规范句子 所有不符合规范句型的句子均为非规范句子。

由于句子译前处理的主要目的是便于机器翻译机制理解原文，以实现正确的翻译，因此，本系统中所有不符合书面语表达习惯，或结构上正确而无实际意义以及所有可能导致机译机制误译的句子，均被视为非规范句子。

三、非规范句子特点分析

根据非规范句子成因，我们将译前处理的非规范句子归纳为两种类型：(1) 由于讲话者的原因，输入系统的句子本身就是不规范的(假定语音识别和后处理机制对输入句子转换正确)，我们称之为先天性非规范句子；(2) 输入系统的句子本身是规范的，但经语音识别和后处理机制转换后，成为不规范的句子，我们把这一类句子称之为损伤性非规范句子。以下我们详细分析这两种类型句子的特点。

3.1 先天性非规范句子

先天性非规范句子主要是由于人们日常会话的习惯自然形成的，各种现象十分复杂，很难准确地归类，根据本系统中句子处理的需要，我们粗略地将其分为如下 4 种情况：

3.1.1 零句

一个完整的句子有主语谓语两部分，零句没有主语－谓语形式[6]。在本系统中，来自语音处理机制的零句多表现为短语形式，或者一个(字)词，或日常套语，因此，机器翻译机制对于大部分零句的翻译并不困难，例如："好极了"；"真倒霉"；"不准抽烟"等，几乎等同于词与词或短语之间的翻译。但有些零句一般不能被机译机制直接理解，必须进行规范化处理，例如，个别用作熟语的不及物动词零句"不信拉倒"，"死了活该"；在特定语境下使用的用于陈述的零句"(在公共汽车上售票员说)西单下"，"(在餐厅里有人说)热水也行"等。

3.1.2 省略句

省略是现代汉语口语的重要特征，从句子的主谓主要成分，到定语、补语、状语等一般成分，几乎都可以省略，这就给句子分析带来极大困难。请看如下例句：

(1) 笔(既然)丢了就丢啦，有什么关系！　(连词省略)

(2) 小李一阵风跑回家。　(意思：小李象一阵风似的跑回家)　(介词省略)

(3) 明天（是）星期天。　(动词省略)

(4) 有空（的时候）来信。　(表示时间、处所等词语省略)

(5) 开车的（人）是军人。　("的"字短语替代整个名词组合)

(6) 老师管教（得）很严。　(粘着成分脱落)

(7) 他是 一　(意思：他是谁？)　(半截句)

3.1.3 重复现象

造成口语里重复现象的原因很多，有时是因为讲话者思维出现故障而人为地填空，有时因为想的跟不上说的而引起重复，也有时是为了加强语气或使用了发语词、口头语等。重复的内容有的完全一样，有的不完全一样，有的是临时插入的内容，有的却是"嗯嗯呀呀"的搪塞语，这种情况必须分别处理。

例句：(1) 你叫什么名字你？

(2) 你看看你看看，又是磁盘，又是光盘的。

3.1.4 松散结构

口语中的松散结构一般表现为 3 种情况：1) 主谓、连动、兼语、状动、并列、同位等

结构，说话时当中有明显的停顿，并有语气词；2）动宾之间有停顿，且有语气词；3）用慢速语流说话，每个词或短语后面有停顿或加语气词。

例句：(1) 电视呀，冰箱呀，我们都有。

(2) 李老师，谢谢你呀！

(3) 我吧，今天吧，给他打了个电话，他吧，不在家。

结构松散的句子经语音处理模块后，一般被断成两个或几个句子，这样就使有的句子成为规范句子或短语，甚至一个词，有的则成为残缺的非规范句子。

在先天性非规范句子中，除了上述列举的一些情况外，还有一些难以判断的句子，如有的句子在句法上正确，而语义上却不成立，例如："草吃兔子"；有的句子字面上的语义分析不成立，但在特定对话语境下却成立，例如："烟抽我"（实际上是"我抽烟"）。因此，这些句子常常需要人工辅助才能断定。

3.2 损伤性非规范句子特点分析

损伤性非规范句子是由于系统语音信号处理机制和音字转换机制的错误或偏差引起的。根据错误形成原因，损伤性非规范句子大体分为如下两种情况：

3.2.1 由语音信号处理机制导致的错误

由于讲话者所处的语言环境不一定都是纯净的，常常受到噪声等其它因素的干扰，语音信号处理模型也存在一定的误差，因此，语音识别后的拼音流时常含有多余音节或错误音节。作为一个实验系统，我们可以保证讲话者的语言环境基本是纯净的，而暂不考虑噪声的干扰，但是对于非特定人的识别音节要达到 100 %的准确几乎是不可能的，在提供多音节候选的情况下，不可避免地使音字转换机制生成的句子含有一定误差。

3.2.2 由音字转换机制导致的错误

即使语音信号处理机制送出的拼音流是完全正确的，音字转换机制也可能因处理不当而导致非规范句子的产生。这种情况又可细分为 2 种：

(1) 由于拼音流分词错误引起音字转换错误，从而导致非规范句子产生；

(2) 由于同音字（词）识别错误而导致非规范句子。

在语音理解中，一旦音节流切分错误，转换结果是必错无疑的。同音字（词）识别是长期困扰音字转换的关键性难题，在目前技术下同音字（词）误选是难以避免的。

四、规范化处理策略

在本系统中，句子规范化处理的目的是易于机器翻译。因此，句子规范化处理模块采取的策略是，有针对性地检查和校正那些容易引起误解、机器难以理解的句子，而对单词和单字不予处理，通过人机交互方式，由用户最终把关。具体的实现策略如下：

4.1 系统检查与人机交互相结合

汉语口语句子自动检查和规范化处理是一项十分困难的工作，从系统实现的实际情况

出发，我们采取系统检查和自动纠正与人机交互相结合的处理策略。处理顺序是：系统先判断输入句子是否为单个字或词，如果是，则执行下面 4.2 处理策略；否则，判断是否为特定句式，如果是，则执行下面的 4.3 策略；否则，如果是在案非规范句子，则系统自动予以纠正，然后将纠正结果交给用户确认；如果是在案规范句子，则直接交给用户确认，否则，系统给用户指示特别信息并记录相应句子，以备用户根据具体情况扩充非规范句型转换规则库或规范句型模式库。

由计算机自动检查，然后人工核对的处理方案是可行的，大大减轻了人的工作量。

4.2 单个字（词）由用户确认

在系统的音字转换中，单个字(词)的确认非常困难，一个不带声调的单音节可以对应 100 多个汉字，一个双音节对应十几个两字词也是很常见的，因此，在相对孤立的零句中同音字（词）的选择一般是根据字（或词）的概率进行的，在目前系统还不能作远距离对话语境分析和理解的情况下，只能由人工来确认同音字（词）的选择是否正确。由于单字或单词一般不对机器翻译机制构成困难，因此，系统无需作其它处理。

4.3 特定句式针对性处理

许多零句以单词、短语或套语的形式存在，按上述策略就可以处理。但也存在一些机器难以直接理解和翻译的熟语，如“不信拉倒”、“死了活该”等，对这些特殊的句式，本系统设立了专门的对应变换，系统对库中存在的熟语按变换句式进行自动转换。例如：

不信拉倒 → 信不信由你 (believe it or not)

死了活该 → 罪有应得 (deserve the punishment)

转换后的内容一般在词典中都作为固定成语或习惯用语有对应的英文翻译，翻译机制不需要作复杂的句法分析就可以完成自动翻译。

4.4 在案非规范句子自动处理

所谓的在案非规范句子是指那些已经被总结、归纳，并记录在特定知识库中的非规范句子。一般地，在特定知识库中，存放了所有被归纳的非规范句子的句型模板和对应的变换规则，分析句子一旦与库中某一模板相匹配，立刻执行相应的转换。其模板格式为：

$$XP_1(a_{11}; a_{12}; \cdots; a_{1m1})\ XP_2(a_{21}; a_{22}; \cdots; a_{2m2})\ \cdots\ XP_n(a_{n1}; a_{n2}; \cdots; a_{nmn})$$

其中 XP_k ($1 \le k \le n$) 是相应的词性，a_{ij} ($1 \le i \le n, 1 \le j \le m_i$) 是对应单词的语义特征信息，每一个 a_{ij} 可以是多个语义特征的逻辑与，分号表示逻辑或。虚词和某些实词可以不含语义特征直接出现在模板或变换句型中，对应的转换跟在模板之后。例如：

(1) NP_1 (pron) NP_2 (name) → NP_1 是 NP_2

实例：我张明 → 我是张明

(2) NP_1 (pron ； name) NP_2 (place-name) 人 → NP_1 出生在 NP_2

实例：张明上海人 → 张明出生在上海

(3) NP_1 (time) NP_2 (date ； week ； fest) → NP_1 是 NP_2

实例：明天星期天 → 明天是星期天

(4) X Mood-Aux → X ； 句子末尾的语气助词被截掉。

实例：谢谢你呀 → 谢谢你

在案非规范句子被自动变换处理后，由人工确认变换正确与否。

4.5 机器学习与人机交互辅助完成知识库修改

根据系统在运行过程中自动记录的信息，通过人机交互方式对非规范句型模板库和规范化句型模板库进行修改和扩充。

规范句型库中的模板格式与非规范句型的模板格式相同，只是没有后面的转换部。模板格式的存储顺序按对应句型的出现频率排列。句子分析前的分词处理直接采用拼音流分词的结果，这样一方面避免了重复处理，另一方面也便于验证音字转换模块的处理结果。

五、问题讨论与结语

汉语句子规范化处理是一项非常困难的工作，它在汉语语音识别和机器翻译等许多领域都有十分重要的意义。本文所做的工作是对口汉语句子规范化处理研究的初步尝试和探索，若干具体问题仍有待于更深入地研究。本课题正在进行中，下一步的工作侧重于如下几个方面：

- □ 更细致地分析和归纳汉语口语句子的特点，总结信息处理用口汉语规范化句型；
- □ 进一步研究非规范句子中非法成分的推断算法；
- □ 增加语义分析在句型处理中的作用，提高句子分析的正确率；
- □ 在可能情况下实现部分句子的系统自动矫正，尽量减少人工干预时间。

参考文献

[1] 李临定，现代汉语句型，商务印书馆，1986。

[2] 北京语言学院句型研究小组，现代汉语基本句型，世界汉语教学，1989年第11期。

[3] 陈建民，现代汉语句型论，语文出版社，1986。

[4] 罗振声，郑碧霞，汉语句型自动分析和分布统计算法与策略的研究，中文信息学报,Vol. 8, No. 2, 1994。

[5] 唐泓英、姚天顺，王宝库，关于汉语句型，中文信息学报，Vol. 7, No. 1, 1993。

[6] 陈建民，汉语口语，北京出版社，1984，pp 88 - 190。

[7] 朱学锋，俞士汶，自动翻译电话与口语信息处理研究，人工智能新进展，清华大学出版社，1994。

[8] 慕勇，孙才，罗振声，汉语文本自动查错与确认纠错系统的研究，计算语言学进展与应用，清华大学出版社，1995。