基于双语语料库和规则库的德汉复合句的转换生成

邸海燕　　柴佩琪　　许玉祥

(同济大学计算机系　上海　200092)

摘要：本文在尽可能全面考虑各种常见语法现象的基础上，总结了德汉语转换生成的规律，提出了基于双语语料库和转换规则库的转换生成方法，并详细讨论了双语语料库和转换规则库的设计及数据库实现。

关键字：　转换和生成　双语语料库　规则库

The Conversion and Generation between German and Chinese Clauses Based on Corpus and Rule Base

Di Haiyan　　Chai Peiqi　　Xu Yuxiang

(Dept. of Computer Science, Tongji Uni. Shanghai, 200092)

ABSTRACT: The conversion and generation rules are summarized based on studying the syntax characteristics of German and Chinese clauses. The design and implementation of corpus and rule base are discussed in detail.

Keywords: Conversion and Generation, Corpus, Rule Base

一、引言

目前，语料库的建设和语料库语言学的崛起，给计算语言学带来了巨大的变化，意味着计算语言学进入了一个崭新的时代。自1989年以来，机器翻译进入一个新纪元，出现了第三代机器翻译系统，其主要特点是引入了语料库的概念。这给了我们一个极大的启发。在我们的德汉机器翻译系统中，开始尝试在德汉复合句的转换和生成中引入语料库的概念。

在进行德汉机器翻译的研究中，我们收集了大量的德语和汉语语料，通过分析和整理，把原始的粗语料加工成精语料库，总结出德语复合句的特点及其德汉语之间的转换规律。我们认为，德语简单句的转换生成适合采用动词配价的方法，而在德汉复合句的转换和生成中同时使用双语语料库和规则库将是一个比较好的方法。

二、双语语料库的设计

德语复合句中有一类句子，通常是一些固定用法及常用句型，它们具有下列特点：

1.德语句型模式和连词位置较固定(一般在分句句首).

2.连词词义唯一，其汉语表达方式也比较固定．例如:

(1)并列连词denn(因为)引导的复合句:

Kupfer ist ein Leiter, denn es leitet den Strom. 铜是导体，因为它能传导电流。

连词denn总是位于后一分句的句首，表示原因，可译为“因为”．

(2)并列连词allein(只是)引导的复合句:

Er wollte die aufgabe vollenden, allein es war zu spät. 他本想完成作业，只是时间太晚了。

连词allein位于后一分句的句首，表示转折，可译为“只是”．

(3)从属连词so dass(以至于)引导的复合句:

Zwischen Brinellhärte und Zugfestigkeit besteht der empegische Beziehung, so dass man aus der Härte auf die Festigkeit schliessen kann.

布氏硬度与拉伸强度之间存在着经验关系，以至于从硬度可以推算出强度来。

连词so dass总是位于后一分句的句首，表示结果，可译为“以至于”．

(4)从属连词nachdem(...以后)引导的复合句有两种形式:

a)nachdem在前一分句句首。

Nachdem die Studenten gefrühstückt haben,gehen sie zur Vorlesung.

大学生们吃完饭以后,就去上课.

b)nachdem在后一分句句首。

Er gab die Tat erst zu,nachdem man ihn überführt hatte.

人家证实他的罪行之后,他才招认.

但每种形式的汉语表达都是唯一的.

有上述特点的连词还有bevor、ehe、seit、seitdem、damit、je... desto...、deshalb等。

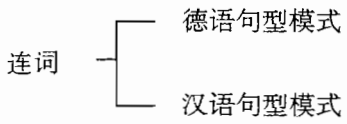
另外，在复合句中有许多常用句型，如常用的带主语从句的结构:

1.Es ist (un)klar, dass...	十分清楚(不清楚)，...
2.Es ist (un)bekannt, dass...	大家知道(不知道)，...
3.Es ist (un)möglich, dass...	...是可能(不可能)的
4.Es ist üblich, dass...	通常，...
5.Es ist wichtig, dass...	重要的是...
6.Es ist notwendig,das...	人们必须
7.Es ist gut(besser),dass...	最好是...
8.Es ist schade,dass...	遗憾的是...
9.Es ist (un)gewiss,dass...	一定（不一定）
10.Es ist (un)wahrscheinlich,dass...	可能（不可能）
11.Es ist wahr,dass...	真实的（不真实的）
12.Es gelingt,dass...	是成功的
13.Es erweist sich,dass...	证明
14.Es ergibt sich,dass...	人们得出（结论）
15.Es stellt sich heraus,dass...	证实
16.Es kommt noch hinzu,dass...	此外，...

17.Es zeigt sich,dass... 表明

18.Es ist zu betonen,dass... 必须强调，...

对于上述句型，我们采用基于连词的双语语料库实现转换和生成，所谓双语语料库是指不需经过语法分析，而是通过德语句子模式直接得到汉语表达模式用于转换和生成的库。其结构如下：



在我们的系统中，我们定义一个开放的符号系统来表示德汉句型模式，目前它包括下面几个符号：

S: 表示一个简单句

Si: 表示第i个简单句(i=1,2,3...)

JVERi: 表示第i个简单句中的比较级

这些元符号是在我们总结出的现有双语语料上提出的。随着新语料的加入，还可以再定义新的元符号。

如deshalb的双语语料表示为：

deshalb S1, deshalb S2.

S1, 所以 S2.

其中，S表示一个简单句，S1表示第一个简单句，S2表示第二个简单句。

je..., desto...的双语语料表示为：

je_desto je JVER1 S1, desto JVER2 S2.

S1越JVER1, S2越JVER2.

其中，JVER表示形容词或副词的比较级。

常用主语从句结构 Es ist möglich, dass...在双语语料表示为：

dass Es ist möglich, dass S.

S是可能的。

nachdem的双语语料表示为：

nachdem Nachdem S1,S2.

在 S1 以后,S2.

S1,nachdem S2.

在 S2 以后,S1.

双语语料库的使用具有下述特点：

1.复合句中的一部分不需经过复合句分析就可以转换生成，提高了速度。

2.双语语料库的库结构为开放结构，具有可重组性和可扩充性，增、删、改十分方便。

3.使生成的汉语不太生硬，更符合汉语习惯。

三、转换规则库的设计

双语语料库解决了一些固定用法及常用句型。但不能解决所有复合句的转换和生成，德语中还有一些复合句不能简单地根据连词和句型来确定转换和生成规则。

这样的复合句有下面几类:

(1)一种连词引导不同类型的从句，其句型可能一致，但其转换规则不一样。

如连词dass可引导主语从句、表语从句、宾语从句、状语从句等，当dass引导主语从句且从句在主句之前时，其形式与dass引导宾语从句且其从句在主句之后时完全一样:

Dass man das kind vernachlassigt, ist ein verbrechen.

不关心孩子是一种不负责任的行为。

Dass unsere These richtig ist, bestatigt sein vertrag. 他的报告证实我们的论点正确。

但其转换规则不一样，对于dass引导的主语从句，我们的转换规则是: 先转换生成从句，再转换生成主句。对于dass引导的宾语从句则是: 先转换生成主句，再从句。

(2)同一种类型的复合句，其德语句型模式一样，但汉语表达不同。如定语从句，某些定语从句适合放在被修饰词前；另一些则适合表达成与主句并列的句子。例如:

Er hat das Examen mit Ausgezeichnung bestanden,wor ü ber (ü ber das) sich seine Eltern sehr freuten. 他考试成绩优异，他的父母对此非常高兴。

由关系副词woruber引导的定语从句，是一种关系从句，适合表达成与主句并列的句子。

Ein Geschenk,worauf(auf das)ich mich besonders freue,ist ein Buch.

最使我感到高兴的一样礼品是一本书。

这种定语从句就适合直接放在被修饰词前。

(3)连词位置不固定，可放句中或句首。

如连词aber、namlich、zwar等引导的复合句。

(4)同一种类型的复合句，其德语句型模式及引导词也一样，但引导词表达的关系不一样，其引导词的汉译也不同。例如:

Der Arzt operiert,und die Krankenschwestern helfen ihm dabei.

医生做手术，护士在一旁帮忙。

此句中的 und 连接两件不同的事情，表示并列关系。und 不译出。

Öffnen Sie das Batteriefach und vom Gerät abheben.

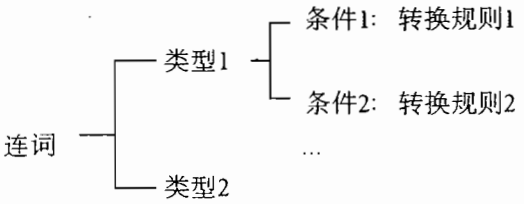
请您打开电池盒盖并且从仪器上拿开。

此句中的 und 连接两个连续的动作，表示递进关系。und 译为“并且”。

对于上述复合句，我们用基于连词的转换规则库进行转换和生成。

依据上述情况，我们先将句子按连词分类。如分成dass类，ob类等。按连词分类后，一部分连词从句即可正确转换。如aber，bald....bald引导的复合句等。在此基础上，对同一连词按类型分类。如dass引导的从句又可分为第一格补足语从句，第四、三、二及介词补足语从句，说明语从句等。按类型分类后，大部分句子均能正确转换。但还有少数情况，在上述两次分类后仍有问题，如定语从句、und引导的并列复合句等，所以还需要其它一定的表层信息及一定的语义信息来解决。

根据上面的分析，我们的转换规则库的结构设计为:



在我们的系统中，我们也定义一个开放的符号系统来表示转换规则，目前它包括下面几个符号:

S: 表示主句

N: 表示从句

C: 表示连接词

SS: 表示主句的主语

NS: 表示从句的主语

C1: 表示连接词汉语意思的前半部分（如果连词的汉语意思是两部分）

C2: 表示连接词汉语意思的后半部分（如果连词的汉语意思是两部分）

DE: 表示汉语的“的”字

P: 表示逗号

E: 表示句号

这些元符号是根据我们总结出的现有规则提出的，随着新情况的出现，还可以再定义新的元符号。

下面以连词 ob 为例说明:

ob 是一个纯连词，它可以引导第一格补足语从句、第四格补足语从句、表语从句等。

1. 引导第一格补足语从句

Ob er morgen kommt.ist noch ungewiss. 是否他明天来还不一定。

ob 在引导第一格补足语从句时，其德汉转换规则是: 先从句，后主句，且连词 ob 的意义放在从句之首。

2. 引导第四格补足语从句

Ich weiss nicht,ob er schon gekommen ist. 我不知道是否他已经来了。

ob在引导第四格补足语从句时，其德汉转换规则是: 先主句，后从句，且连词ob的意义放在从句之首。

3. 引导表语从句

Die Frage ist,ob diese Funktion eine gerade Funktion ist.问题是是否此函数是一个线性函数。

ob 在引导表语从句时，其德汉转换规则是: 先主句，后从句，且连词 ob 的意义放在从句之首。

由此可总结出ob的转换规则，它在转换规则库中的表述为:

ob	第一格补足语从句	NULL	C N S
	第四格补足语从句	NULL	S C N
	表语从句	NULL	S C N

四、双语语料库和转换规则库的数据库实现

虽然目前我们的语料库和规则库的规模还不大，但是由于我们的语料库和规则库具有开放结构，所以，随着德汉机译系统的日益发展和完善，语料库和规则库的规模会日益增大，从而使得它的管理变得繁杂、困难。鉴于此，我们采用Foxpro2.0数据库管理系统软件来建立并管理语料库和规则库。

双语语料库的数据库结构为:

字段名	类型	长度	
conj	字符型	31	: 连词
Gpattern	字符型	50	: 德语模式
Cpattern	字符型	50	: 汉语模式

数据库中有三个字段，每个字段都是字符型。

部分数据库记录示例如下:

RECORD	Gpattern	Cpattern	conj
1	Es ist klar,dass S	十分清楚，S	dass
2	Es ist möglich,dass S	S 是可能的。	dass
...	...	...	...

规则库的数据库结构为:

字段名	类型	长度
conj	字符型	31
type	字符型	20
rule	字符型	50

此数据库有三个字段，其中conj表示连词，type表示从句类型,rule表示转换规则。

部分数据库记录示例如下:

RECORD	conj	TYPE	rule
1	ob	Subjectsatz	C N S
2	ob	Objectsatz	S C N
3	ob	Linksatz	S C N
...	...	...	...

五、结束语

我们在进行德汉机器翻译的研究过程中，对德汉复合句的类型和特点做了比较深入的研究，提出了基于双语语料库和规则库的德汉转换生成方法。这在机器翻译的领域中还只是尝试，有待于进一步实验和完善。

参 考 文 献

【1】韩万衡编 《德语配价语法》 商务印书馆

【2】俞士汶 “自然语言的歧义与机器翻译的对策” 中文信息学报Vol.3 No.2

【3】周明 黄昌宁等 “统计与规则并举的汉语句法分析模型” 计算机研究与发展Vo,.31 No.2

【4】刘开瑛、郭炳炎 《自然语言处理》 科学出版社 1991

【5】柴佩琪等 “基于动词配价的德汉机器翻译” 《1992 全国机器翻译学术会议论文》