文章编号：1003-0077(2008)02-0047-08

基于派生文法的日—蒙动词短语机器翻译研究

百 顺

（筑波大学 大学院图书馆情报媒体研究科 日本 筑波市 305-8550）

摘 要：本文探索了源语为日语，目标语为蒙古语的动词短语机器翻译系统的实现方式。基于主张日语不活用的派生文法，重新分析日语附加成分。将日语的词干和附加成分转换到蒙古语的词干和附加成分之后，运用蒙古语的语音规则来处理并生成动词短语。在此基础上试做了日—蒙动词短语机器翻译系统。对 30 篇日文报道的 403 个动词短语进行测试，取得了 95.78% 的正确率。

关键词：人工智能；机器翻译；派生文法；日语附加成分的分析；语音规则；短语生成

中图分类号：TP391　　**文献标识码**：A

Research on Japanese-Mongolian Machine Translation of Verb Phrase Based on the Derivational Grammar

BAI Shun

(Graduate School of Library, Information and Media Studies
University of Tsukuba, Tsukuba 305-8550, Japan)

Abstract: This paper describes the implementation of a Japanese-Mongolian verbal phrase machine translation system of verb phroses. In the Japanese derivational grammar, there is no concept of conjugations, a word is analyzed into stems and suffixes. After translating Japanese stems and suffixes into Mongolian stems and suffixes, Mongolian phonetic rules are used to process and generate verbal phrases. We implemented a Japanese-Mongolian verbal phrase machine translation system. We also tested 403verb phrases from 30Japanese reports, and achieved a 95.78% accuracy.

Key words: artificial intelligence; machine translation; derivational grammar; Japanese suffixes analysis; phonetic rule; phrase generation

1 引言

日—蒙机器翻译系统尚未问世。对蒙古语文信息处理来说，从英语以及日语到蒙古语的机器翻译的研究是具有非常重要的理论和实践意义的课题。于是，作者做了以激活日语和蒙古语的类似性，实现日—蒙机器翻译系统为目标的一些尝试性研究。所谓的短语是由两个或两个以上的词或短语按照一定的公式构成的，能够在句子中承担某种功能的语法单位[3]。在本文中描述的动词短语是由动词词干（包括动词性合成词词干）上接加附加成分串而构成的。

蒙古语属于黏着型语言，语法体系和日语有很多相似之处。特别是两种语言的句法，是几乎相同的。因此，对于从日语到蒙古语的翻译来说，把日语语句的分析结果直接翻译为蒙古语，也可以译出一定程度的译文。当然，对日语和蒙古语而言，两种语言之间的语音变化、构词构形和语法等方面也存在很多不同点，也有必要对词干和附加成分的翻译进行适当的择词等处理。

在日语方面运用了派生文法[1]。理由是：(1)日语传统语法的活用形处理是机器翻译中的难点之一。(2)派生文法是基于黏着语性质的语法。它是把日语的构词构形作为词干上接加附加成分描

收稿日期：2007-04-18　**定稿日期**：2007-09-07

作者简介：百顺(1964—)，男，博士生，主要研究方向为机器翻译、自然语言处理。

述的。因此,不需要活用形处理。(3)日语派生文法比传统语法更为接近蒙古语语法。因此,本文利用派生文法对日语词法分析结果中的附加成分重新加以分析,把日语的词干和附加成分转换成蒙古语的词干和附加成分,运用蒙古语语音规则生成蒙古语动词短语,并提出了日一蒙动词短语机器翻译系统的实现手法。

2 基于派生文法的日语动词短语的解析

基于日语作为黏着语的性质,派生文法的观点认为日语没有活用性[1,6]。所谓黏着语,就是其语法机能由附加成分表现出来的语言的总称[1]。它的构词构形是词干上接加不同附加成分来完成的。派生文法对日语动词短语的描述概括起来有下列特征。

2.1 动词短语的形成

派生文法把日语的独立词大体上分为动作动词词干、形状动词词干、实名词词干、形状名词词干等四种。这些分类与日语传统语法中的动词、形容词、名词、形容动词相对应。

派生文法的动词短语是由动作动词词干(一次词干)上接加附加成分串构成的。附加成分,大体上可分为两种:机能附加成分和派生附加成分。

2.2 派生附加成分和机能附加成分

对于派生文法来说,把「書カセル」解析为 kakase-ru,即词干 kak-上接加附加成分-ase-而派生二次词干「書カセ」kakase-。这是动作动词词干上接加附加成分而派生的新的词干。这种附加成分叫做派生附加成分[1]。

对派生附加成分而言,像-ru 这样不派生新词干的附加成分叫做机能附加成分[1]。动作动词词干上接加多个附加成分时,机能附加成分排在最后。

2.3 元音词干和辅音词干

动词的不变化部分,即去掉附加成分之后剩下的部分叫做动词词干。以传统语法的一段动词「起キル」「食ベル」为例,不变化部分「起キ」「食ベ」是动词词干,这些词干都以 i 或 e 来结尾。像这种以元音结尾的动词词干叫做元音词干。以五段活用动词「話ス」为例,在传统语法中词尾变化是「話サ」「話シ」「話ス」「話セ」「話ソ」。从语音学的角度能把这些活用形考虑为「hanas-a」「hanas-i」「hanas-u」「hanas-e」「hanas-o」。其中 hanas 是不变化部分,像这种以辅音结尾的动词词干叫做辅音词干。

2.4 连接辅音和连接元音

动作动词词干上接加附加成分时,必须遵从以下 2 条规则。

规则 1:辅音结尾的词干上接加以辅音为首的附加成分时,附加成分首的辅音要脱落。

规则 2:元音结尾的词干上接加以元音为首的附加成分时,附加成分首的元音要脱落。

规则 1 这种会脱落的辅音叫做连接辅音[1]。例如:辅音词干 hanas 上缀接附加成分 ru 时,附加成分首的辅音 r 就会脱落,变成 hanasu。

规则 2 这种会脱落的元音叫做连接元音[1]。例如:元音词干 tabe 上缀接附加成分 ita 时,附加成分首的元音 i 就会脱落,变成 tabeta。

派生文法中为了表示以上所看到的这些语法现象,必须由音素单位的罗马字来表述。

2.5 词干的词类变化

派生文法所述的是在词干上接加附加成分时会产生词类变化。因此,把词干后接的附加成分看作是有限状态自动机的输入,其词类变化为状态变迁。如图 1 所示的是在派生文法中词干上接加附加成分

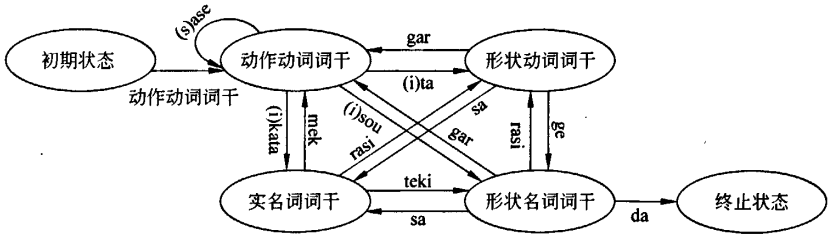

图 1 派生文法的有限状态自动机模型

时所发生的语法机能变化模型①。词干的语法机能的变化就是自动机的状态变迁，这是由于派生附加成分的特性使词干的词类变化所产生的结果。譬如，图 1 中显示的是一次词干的动作动词词干上接加附加成分而出现的状态变迁，也就是说从一种词干变迁到另一种新的词干的例子。

3　用派生文法的动词短语的翻译

日语和蒙古语都具有黏着语的性质，所以两种语言的动词短语的形成具有共同的特点：都是词干上接加附加成分所形成的。例如，和日语词干「食べ」相对应的蒙古语词干是「ide」。表达使役态时，把使役态附加成分「-(s)ase-」接加在词干上，就派生为「食べsase-」。同样，在蒙古语的词干上接加相对应的使动态附加成分「gul」，就派生为「idegul」。同时也存在一些不同点。例如，由于两种语言在敬语方面的表达方式有所不同，表达尊敬的日语附加成分「-(r)are-」和「-(i)mas-」相对应的蒙古语的附加成分就不存在。还有，日语使役态附加成分「-(s)ase-」相对应的蒙古语的使动态附加成分有「Gol，gul」「lGa，lgc」「Ga，ge」等三组。现阶段，本系统只限于第一组「Gol，gul」。为了便于处理，以后把「Gol，gul」等表达同样的语法意义并且相对立的这种附加成分记为「[Gg][ou]l」。下面把日语动词短语翻译为蒙古语的过程表示为图 2。

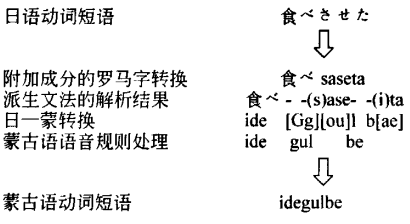

图 2　基于派生文法的动词短语翻译例

4　翻译过程中存在的难点

日语和蒙古语虽然在句法和构词方面有很多相似之处，但也有不同点。例如，否定性动词短语中的附加成分的结构不同。还有，日语和蒙古语不能一一对应的现象。在这种情况下生成的动词短语就成为不正确的或不自然的译文。在本章里，将对这些问题举例说明。

4.1　同形附加成分的区分

日语在派生文法中存在同形附加成分，但其对应的蒙古语的附加成分就不同。例如，表达愿望的附加成分和表达过去式的附加成分都是「-(i)ta」。而且，对同一个表达过去式的附加成分来说，它的连体形和终止形也是「-(i)ta」。但其蒙古语的译文是不相同的。如图 3 所示。

图 3　同形附加成分相对应的蒙古语附加成分例

以上三种附加成分的形状都是「-(i)ta」，但其在蒙古语中，表达愿望的附加成分相对应的附加成分为「ye」，表达过去式附加成分的连体形为「gsen」，表达过去式附加成分的终止形为「be」。因此，要翻译日语附加成分「-(i)ta」时，必须根据其机能和动词形，从「ye」，「gsen」和「be」中进行适当的选择。

4.2　サ变名词和サ变动词的翻译问题

在日语里，像「勉強する」「感謝する」似的，存在一些名词上后接「～する」的词。这类名词叫做サ变名词。名词后接「～する」，把名词动词化的词，在语法上叫做サ变动词。サ变名词通常指的是表示动作的名词。和サ变动词一起，作为合成词来使用。日语里，サ变名词和サ变动词的数量很多。

在蒙古语中，和「～する」相当的单词是「hihu」。譬如，日语的「勉強」相对应的动词是「sorolqa」。作为合成词来使用的「勉強する」要译成蒙文时，如果把「勉強」和「～する」直接翻译的话，就成为两个动词「sorolqa」「hihu」。这样的译文是不正确或不自然的。其实蒙文里有「sorolqaho」这样的译文。于是，要正确翻译「勉強」「～する」，必须使它变为「勉強す＋(r)u」这样一个词干上接加附加成分的形态。

4.3　语序的不一致

日语和蒙古语语序虽然有很多相似之处，但也有不同点。例如，表达否定性动词短语的过去式时，两种语言的附加成分的顺序就会有不同之处。表达现在和未来式的时候，蒙古语有必要追加附加成分。

① 这是作者根据派生文法研究出来的有限状态自动机模型。

下面以「食べnakatta」和「食べnai」两个动词短语为例，看看两种语言之间的不同点。

例1：食べnakatta

日语： 动词词干 否定式附加成分

蒙古语： 动词词干 过去式附加成分

日语： 过去式附加成分

蒙古语： 否定式附加成分

例2：食べnai

日语： 动词词干

蒙古语： 动词词干 非过去式附加成分

日语： 否定式附加成分 附加成分

蒙古语： 否定式附加成分 附加成分

例1表示：日语附加成分的顺序是，否定式附加成分在前而过去式附加成分在后；蒙古语附加成分的顺序是，过去式附加成分在前而否定性附加成分在后。例2表示：日语否定性附加成分直接和词干连接，而蒙古语的词干和否定性附加成分之间必须要追加非过去式附加成分。有一些日语附加成分没有相对应的译文，譬如本例中的「i」就是其中的一个。

5 对难点问题的解决

本文不仅要利用语言之间的类似性，而且要进行句法分析。下面要论述对第四章里提出的疑难问题的解决方法。

5.1 同形附加成分的区分

同形附加成分的区分是要根据句法单位内的附加成分的黏着前状态和黏着后状态来决定的。所谓的附加成分的黏着前状态就是黏着什么样的词干。所谓附加成分的黏着后状态就是黏着之后派生什么样的二次词干或者形成动词形的连用形、连体形、终止形和命令形的哪一种。譬如对附加成分「(i)ta」来说，在表示愿望和表示过去式的场合，虽然黏着前状态都是动词词干，但是黏着后状态就不同。前者的黏着后状态是形状动词词干，后者是动作动词连体形和终止形。这种不同状态能使同形附加成分有区分开来的可能性。因此，本系统解决了对同形附加成分的区分问题，也实现了图3中的表示愿望的「(i)ta」译为「ye」，表示过去式附加成分「(i)ta」的连体形译为「gsen」，终止形译为「be」。

5.2 对サ变名词和サ变动词的翻译问题的处理

关于第四章里提出的サ变名词和サ变动词翻译问题的对策是把同一个句法单位的サ变名词和サ变动词用以下规则来合成一个动作动词。

规则：サ变名词 + サ变动词 → 动作动词

例如，把「勉強」和「する」合并为「勉強する」。因此，把词干部分「勉強す」和附加成分「(r)u」分别译为「sorolqa」和「ho」。这样就生成了「sorolqaho」的很自然的蒙古语译文。

5.3 语序不一致的调整

本文从派生文法的角度把日语句法单位看作是词干上接加附加成分串的形式。关于第四章第3节里举的否定性动词短语的问题，对附加成分的顺序制作了调整规则，运用这些规则对附加成分的顺序进行处理。以下表示的是具体的规则。

规则1：动作动词词干 + 否定式附加成分 + 过去式附加成分 →
动作动词词干 + 过去式附加成分 + 否定式附加成分

规则2：动作动词词干 + 否定式附加成分 →
动作动词词干 + 非过去式附加成分 + 否定式附加成分

运用以上规则解决了在第四章第3节中举的例1例2的附加成分的调整问题。把日语的词干和附加成分转换成蒙古语的词干和附加成分，用语音规则生成动词短语。结果是：把「食べnakatta」译为「idegsen ugei」，把「食べnai」译为「idehu ugei」的很自然的译文。

6 机器翻译系统的实现

6.1 系统的构造

本系统由四个部分组成(图4)。也就是词法句法分析，基于派生文法的附加成分分析，日语—蒙古语转换和蒙古语短语生成等。

词法分析利用了日语词法分析系统JUMAN，句法分析利用了日语句法分析系统KNP。

对于KNP分析出来的短语进行基于派生文法的附加成分分析和蒙古语语音规则处理。

6.2 基于派生文法的附加成分分析

本模块是由5个部分模块组成（图5)。

6.2.1 词干整理

派生文法基于日语作为黏着语的性质，认为词

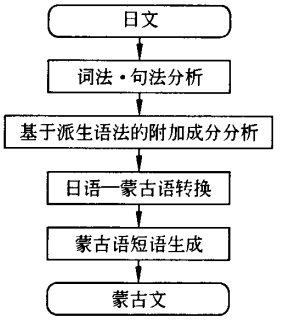

图 4　日—蒙机器翻译系统流程图

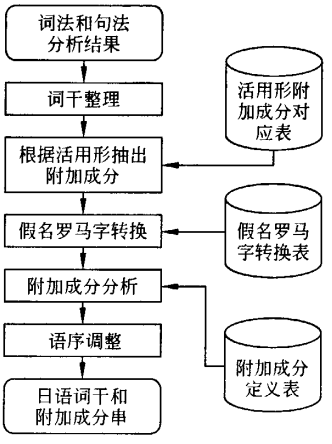

图 5　附加成分分析模块

干上接加附加成分而构词或构形。词干整理基于这个观点，把 KNP 分析结果的短语当作一个词干和附加成分的连接形式。但是，日语句法分析系统 KNP 分析出来的短语里有可能存在多个独立词的情况。针对这个问题，有必要把这些多个独立词并为一个合成词来处理。例如，在「一括処理する」的短语里包含「一括」、「処理」、「する」等三个独立词。把这三个独立词并为一个「一括処理する」的合成词。合成词的词类由最后的独立词的词类来决定。

6.2.2　根据活用形抽出附加成分

派生文法认为，日语不存在活用。也就是说，动词（含形容词、形容动词）是在词干上接加附加成分而构成的，因此存在着构成各种活用形的机能附加成分。例如，表 1 所表示的是构成辅音动词「書く」的活用形的附加成分。

根据活用形抽出附加成分的目的是要抽出表 1 所表示的那种构成动词（含形容词、形容动词）的活用形的附加成分。方法是：利用 JUMAN 分析出来

表 1　活用形形成附加成分的例子

活 用 形	传统文法	派生文法	附加成分
未然形	書か	kak-a	a
基本连用形	書き	kak-i	i
基本形	書く	kak-u	u
基本条件形	書けば	kak-eba	eba
意志形	書こう	kak-ou	ou
命令形	書け	kak-e	e

表 2　活用形附加成分对应表的一部分

活用形—活用形	附加成分
辅音动词カ行—未然形	a
辅音动词カ行—基本形	u
辅音动词カ行—意志形	ou
辅音动词カ行—命令形	e
辅音动词カ行—基本条件形	eba
辅音动词カ行—基本连用形	i

的有活用形的独立词（动词、形容词、形容动词）的活用形和活用形的信息，作了附加成分表。根据这个表，进行构成活用形的附加成分的抽出。表 2 所表示的是由于活用型和活用形而得出的附加成分表的一部分。

6.2.3　附属词的罗马字转换

派生文法是以语音学、形态学为中心的文法。把附加成分作为语音单位来考虑。因此必须用罗马字来表示。附加成分的罗马字转换正是从这个观点出发的。因此，作者制作了日语的假名和罗马字对应表，根据这个表，对 JUMAN 的词法分析出来的附加成分进行罗马字转换处理。表 3 表示的是假名和罗马字对应表的一部分。

表 3　假名和罗马字对应表的一部分

假名	せ	ら	れ	た
罗马字	se	ra	re	ta

6.2.4　附加成分的分析

根据附加成分定义表，对词干上接加的附加成分串进行重新分析。派生文法是正规文法，因此可以作为有限状态自动机来处理（图 1）。根据附加成分定义表，把自动机的状态作为词干种类，输入为附加成分。因此，这个表具有使附加成分模型化的意义。定义表记录了派生文法中包含的全部附加成

分。具体的分析方法是这个自动机接受正规表达式对附加成分的匹配。附加成分定义表的一部分为表4所示。

6.2.5 附加成分的顺序调整

附加成分顺序的调整在第5章第3节中说明过，此处不再赘述。

表4 附加成分定义表的一部分

黏着前状态	语法机能	黏着后状态	附加成分
动作动词词干	非完成态肯定	动作动词连体形	(r)u
动作动词词干	完成态肯定	动作动词连体形	(i)ta
动作动词词干	前望态肯定	动作动词连体形	(y)ou
动作动词词干	使役	动作动词词干	(s)ase
动作动词词干	被动	动作动词词干	(r)are
形状动词词干	非完成态	形状动词终止形	i
形状动词词干	完成态	形状动词终止形	katta

6.3 日—蒙转换模块

利用日—蒙词干词典和日—蒙附加成分对照表，把日语的词干和附加成分分别译为蒙古语的词干和附加成分。

日蒙附加成分对照表包括：派生文法的全部附加成分和根据翻译系统的需要而补充的一些附加成分。在蒙古语附加成分里有很多同音异形附加成分，因此，蒙古语附加成分的译文里不只是一对一，还有一对多的情况。例如，日语属格助词「の」相对应的蒙古语附加成分是 yin，on，un，o，u 等五种。

6.4 蒙古语生成模块

运用语音规则把蒙古语词干和附加成分连接起来，生成蒙古语短语。蒙古语语音规则是根据蒙古语语法[2]，又从符合自然语言处理的角度制作的。蒙古语语音规则主要包括元音和谐规则、元音和辅音相连规则、辅音和谐规则、连接元音书写规则等。图6表示的是应用语音规则生成蒙古语短语的例子。

日语短语	食べsaseta
附加成分分析	食べ--(s)ase--(i)ta
日—蒙转换	ide [Gg][ou]l [Gg]s[ae]
元音和谐规则	ide [Gg]ul [Gg]sen
辅音和谐规则	
元音辅音结合规则	ide gul gsen
连接元音书写规则	ide gul、u gsen
蒙古语短语生成	idegulugsen

图6 应用语音规则生成蒙古语短语的例子

7 实验评价

在本章里，运用作者试作的翻译系统进行了动词短语的实验，并做出评价。

7.1 对象数据

为了试验本系统的翻译精度，在日本每日报[9]的有关农、林、牧、水的310个新闻报道中，用机械选择了30个新闻报道作为测试数据库。其余的280个新闻报道作为训练数据库。而且以测试数据库作为对象，用本系统进行了翻译，对其生成的403个不同动词短语进行了评价。

7.2 评价方法

评价是由作者以外的一位蒙古族人来实施的。在这里，所谓的正确译文就是作为蒙古语完全正确的表述。所谓的错误译文就是语法或意义上不正确的表述。这次评价因为不是整个句子的评价，而是以短语为单位，并且重点放在附加成分的分析和蒙古语语音规则的准确率上，所以对词典里没有记录的单词，进行了一些适当的补充。

7.3 结果和考察

针对测试数据库30篇新闻报道进行翻译的结果生成了481个蒙古语动词短语。其中有一些重复的，不同动词短语的数量为403个，正确翻译的动词短语有386个，获得了95.78%的正确率(表5)。

表 5　正确翻译率

不同动词短语数	正确翻译数	正确翻译率
403	386	95.78%

下面把错误译文的详细原因表示为表 6。

表 6　错误翻译的原因细目

错误翻译的原因	个数	错误翻译率
同形附加成分的区分	0	
サ变名词和サ变动词处	0	
语序的不一致	0	
附加成分分析失败	1	0.25%
语音规则处理	13	3.23%
多义词	3	0.74%
合　计	17	4.22%

表 6 当中的同形附加成分的区分，サ变名词和サ变动词的处理以及语序的不一致是在第四章里谈到的问题。这次试验当中同形附加成分的区分问题出现的次数为 103 次，其中表示愿望的场合是 1 次，表示过去式连体形的场合是 21 次，表示过去式终止形的场合是 81 次。由于本系统的特殊处理而全部翻译为正确译文。

关于サ变名词和サ变动词的处理(85 次)和语序的不一致(16 次)问题，由于本系统中采取适当的对应措施而翻译的译文也是全部正确的。

关于附加成分的分析是基于派生文法的最关键的环节。在文献[5]里附加成分分析的失败占错误翻译率的 85%，通过附加成分定义规则的强化，这次试验中失败的个数是 1 个。不过，这也是在含有文言文的动词短语的场合出现的失败。

这次试验中需要语音规则处理的地方有 569 个。词干和附加成分，附加成分和附加成分之间的连接处都需要语音规则的处理。所以，语音规则处理的个数比动词短语要多得多。语音规则处理的总数 569 次中，由语音规则而引起的失败个数是 2 个，由补助动词的处理而导致的失败个数是 11 个。

蒙古语语音规则里有一些特殊现象，要对这些特殊现象进行处理，现在的语音规则还不够充分，还需要探讨和强化。还有，在本文中是把补助动词作为附加成分来处理的。这次试验中补助动词语音处理的失败次数最多。譬如，本系统把「生産している」翻译为「uiledburileju bain-e」，这是错误译文。正确译文应该是「uiledburileju bain-a」。这是因为，现系统中日语动词短语和蒙古语动词短语都是由词干上接加附加成分串构成的。特别是蒙古语的语音处理是由词干(一次词干)的性质决定附加成分的性质，并且选择符合一次词干性质的附加成分。例子中的一次词干「uile」是阴性词干。按元音和谐规则，一次词干(阴性)上接加的附加成分都是阴性的。但是，蒙古语的补助动词「bain-a」是个独立词。也是词干上后接附加成分构成的。因此，按理说是补助动词的词干「bai」决定其后接加的附加成分的性质。但是，现在的系统是补助动词词干上接加的附加成分也受一次词干「uile」的制约，所以导致语音规则处理的失败。对补助动词的语音规则处理问题作为今后的研究课题保留。

多义词处理是整个机器翻译的难点之一。对于日—蒙机器翻译来说，多义词包括词干的多义和附加成分的多义两种。对附加成分的多义而言，日语动词构词构形附加成分的多义词不多。其中，在本文里对某些多义附加成分进行了处理。比如说，在第五章第 1 节里解决的同形附加成分「(i)ta」的区分就是在某种意义上的多义词处理。是根据词干和附加成分的连接前状态和连接后状态处理的。在本文中，对词干的多义词还没有进行处理。譬如，本系统把「あるとき」翻译为「bain-a qaG」，这是错误译文。正确译文应该是「jarim uy-e」。这次试验当中出现了向这种不符合原文意思的译文只 3 个。这是因为，其一，日语和蒙古语有很多相似之处。其二，本系统训练数据库和测试数据库都是关于农、林、牧、水的新闻报道。为了提高本系统的正确翻译率，对多义词的处理作为下一个阶段的研究目标。

8　结论

本文以日语到蒙古语的机器翻译系统的开发为目标，提出并实现了基于日语派生文法的动词短语的翻译方式。根据试验结果，证明了动词短语的高精度的翻译是可能的。

今后，为了提高动词短语的正确翻译率，强化对补助动词的处理。研究的重点放在多义词的处理并实现具有实用性的日—蒙机器翻译系统。

致谢　日本东京大学石川徹也特任教授和筑波大学长谷部纪元教授对本研究给予了极大的支持和精心的指导。在此表示衷心的感谢。

参考文献：

[1] 清濑义三郎则府. 日本語文法新論—派生文法序説[M]. 东京:桜楓社,1989.

[2] 清格尔泰. 蒙古语语法[M]. 呼和浩特:内蒙古人民出版社,1991.

[3] 巴达玛敖德斯尔. 面向机器翻译的汉蒙短语转换规则研究[M]. 呼和浩特:内蒙古教育出版社,2005.

[4] 那顺乌日图,刘群,巴达玛敖德斯尔. 关于汉蒙机器辅助翻译系统[J]. ALTAI HAKPO, 2001,(11): 35-41.

[5] 百順,長谷部紀元,石川徹也. 派生文法基づく日本語からモンゴル語への文節翻訳[A]. 言語処理学会第12回年次大会発表論文集[C]. 东京: 2006,584-587.

[6] 小川泰弘,ムフタル・マフスット,杉野花津江,稲垣康善. 派生文法による日本語形態素解析[A]. 情報処理学会論文誌[C]. 1999,40(3):1081-1090.

[7] 伊·达瓦,张玉洁等. 蒙古语语言一文字的自动化处理[J]. 中文信息学报. 2006,20(4):56-62.

[8] 聂建云,陈江. 利用平行网页建立中英文统计翻译模型[J]. 中文信息学报, 2001,15(1):1-12.

[9] CD一每日新聞. 東京:每日新聞社,2002.

(上接第27页)

表2 结果分析对比表

	待消解对的总数目	模型识别出待消解对总数目	正确识别出的待消解对数目	准确率	召回率	F值
人称代词消解	124	87	54	62.07%	49.07%	54.81%
互为别名或简称的消解	42	29	15	51.72%	37.8%	43.68%
指示代词消解	256	142	42	29.58%	18.98%	23.12%
本实验的共指消解	379	265	208	78.49%	54.74%	64.5%

参考文献：

[1] 王厚峰. 指代消解的基本方法和实现技术[J]. 中文信息学报,2002,16(6): 9-17.

[2] 王厚峰,何婷婷. 汉语中人称代词的消解研究[J]. 计算机学报,2001,24(2): 136-143.

[3] 李国臣,罗云飞. 采用优先选择策略的中文人称代词的指代消解[J]. 中文信息学报,2005,19(4): 24-30.

[4] 许敏,王能忠,马彦华. 汉语中指代问题的研究及讨论[J]. 西南师范大学学报,1999,24(6): 633-637.

[5] 钱伟,郭以昆,周雅倩,吴立德. 基于最大熵模型的英文名词短语指代消解[J]. 计算机研究与发展,2003,40(9): 1337-1343.

[6] Wee Meng Soon, Hwee Tou Ng t, Daniel Chung Yong Lim. A Machine Learning Approach to Coreference Resolution of Noun Phrases[J]. Computational Linguistics,2001,27(4): 521-544.

[7] 刘群. 统计机器翻译综述[J]. 中文信息学报. 2003, 17(4): 1-12.

[8] M Vilain, J Aberdeen et al, A model-theoretic coreference scoring scheme[A], Proc. Of the 6th Message Understanding Conf (MUC6) [C]. 1995. 45-52.

[9] A. Berger, V. Della Pietra, S. Della Pietra. A Maximum Entropy Approach to Natural Language Processing[J]. Computational Linguistics, 1996, 22(1): 39-71.

基于派生文法的日-蒙动词短语机器翻译研究

作者： 百顺， BAI Shun
作者单位： 筑波大学,大学院阿书馆情报媒体研究科,日本,筑波市,305-8550
刊名： 中文信息学报 ISTIC PKU
英文刊名： JOURNAL OF CHINESE INFORMATION PROCESSING
年，卷(期)： 2008，22(2)
引用次数： 0次

参考文献(10条)

1. 清瀬義三郎則府 日本語文法新論-派生文法序說 1989

2. 清格尔泰 蒙古语语法 1991

3. 巴达玛敖德斯尔 面向机器翻译的汉蒙短语转换规则研究 2005

4. 那顺乌日图. 刘群. 巴达玛敖德斯尔 关于汉蒙机器辅助翻译系统 2001(11)

5. 百顺. 長谷部紀元. 石川徹也 派生文法基づく日本語からモンゴル語への文節翻訳 2006

6. 小川泰弘. ムフタル・マフスット卜. 杉野花津江. 稲垣康善 派生文法による日本語形態素解析 1999

7. 伊・达瓦. 张玉洁. 上园一知. 大川茂树. 章森. 井佐原均. 白井克彦 蒙古语语言-文字的自动化处理[期刊论文]-中文信息学报 2006(4)

8. 聂建云. 陈江 利用平行网页建立中英文统-计翻译模型[期刊论文]-中文信息学报 2001(1)

9. CD-每日新聞 2002

10. 这是作者根据派生文法研究出来的有限状态自动机模型

相似文献(10条)

1. 期刊论文 熊德意. 刘群. 林守勋. XIONG De-yi. LIU Qun. LIN Shou-xun 基于句法的统计机器翻译综述 -中文信息学报2008,22(2)

本文对基于句法的统计机器翻译进行了综述.按照模型所基于的语法不同,将基于句法的统计机器翻译分为两大类:基于形式化语法和基于语言学语法.对这两个不同类别,我们分别介绍它们代表性的工作,包括模型的构建、训练和解码器的设计等,并对比了各个模型的优点和缺点.最后我们对基于句法的统计机器翻译进行了总结,指出设计句法模型时要注意的问题,并对未来的发展趋势进行了预测.

2. 期刊论文 徐波. 史晓东. 刘群. 宗成庆. 庞薇. 陈振标. 杨振东. 魏玮. 杜金华. 陈毅东. 刘洋. 熊德意. 侯宏旭. 何中军. XU Bo. SHI Xiao-dong. LIU Qun. ZONG Cheng-qing. PANG Wei. CHEN Zhen-biao. YANG Zhen-dong. WEI Wei. DU Jin-hua. CHEN Yi-dong. LIU Yang. XIONG De-yi. HOU Hong-xu. HE Zhong-jun 2005统计机器翻译研讨班研究报告 -中文信息学报2006,20(5)

2005年7月13日至15日,中国科学院自动化研究所、计算技术研究所和厦门大学计算机系联合举办了我国首届统计机器翻译研讨班.本文主要介绍本次研讨班参加单位的测试系统和实验结果,并给出相应的分析.测试结果表明,我国的统计机器翻译研究起步虽晚,但已有快速进展,参评系统在短期内得到了较好的翻译质量,与往年参加863评测的基于规则方法的系统相比性能虽还有差距,但差距已经不大.从目前国际统计机器翻译研究的现状和发展趋势来看,随着数据资源规模的不断扩大和计算机性能的迅速提高,统计机器翻译还有很大的发展空间.在未来几年内,在基于短语的主流统计翻译方法中融入句法、语义信息,必将成为机器翻译发展的趋势.

3. 期刊论文 杜伟. 陈群秀. DU Wei. CHEN Qun-xiu 多策略汉日机器翻译系统中的核心技术研究 -中文信息学报2008,22(5)

多策略的机器翻译是当今机器翻译系统的一个发展方向.该文论述了一个多策略的汉日机器翻译系统中各翻译核心子系统所使用的核心技术和算法,其中包含了使用词法分析、句法分析和语义角色标注的汉语分析子系统,利用双重索引技术的基于翻译记忆技术的机器翻译子系统、以句法树片段为模板的基于实例模式的机器翻译子系统以及综合了配价模式和断段分析的机器翻译子系统.翻译记忆子系统的测试结果表明其具有高效的特性;实例模式子系统在1 559个句子的封闭测试中达到99%的准确率,在1 500个句子的开放测试中达到85%的准确率;配价模式子系统在·3 059个句子的测试中达到了89%的准确率.

4. 期刊论文 孙连恒. 杨莹. 姚天顺 OpenE:一种基于n-gram共现的自动机器翻译评测方法 -中文信息学报2004,18(2)

在机器翻译研究领域中,评测工作发挥着重要的作用,它不仅仅是简单地对各个系统输出结果进行比较,它还对关键技术的发展起到了促进作用.译文质量的评测工作长期以来一直以人工的方式进行.随着机器翻译研究发展的需要,自动的译文评测研究已经成为机器翻译研究中的一个重要课题.本文讨论了基于n-gram共现的自动机器翻译评测框架,介绍了BLEU、NIST、OpenE三种自动评价方法,并通过实验详细分析了三种方法的优缺点.其中的OpenE采用了本文提出了一种新的片断信息量计算方法.它有效地利用了一个局部语料库(参考译文库)和全局语料库(目标语句子库).实验结果表明这种方法对于机器翻译评价来说是比较有效的.

5. 学位论文 李剑 英汉机器翻译中的句型转换和译文生成 2005

随着对外交流的日益广泛，机器翻译的研究与实现有着重要的现实意义。同时，机器翻译的研究对于自然语言理解、人工智能、计算语言学等学科的研究也起着重要的推动作用，并对促进情报获取工作发展具有重要的意义。 机器翻译(MT)就是应用计算机实现从一种自然语言文本到另一种自

然语言文本的翻译。20世纪90年代以来，机器翻译的方法基本上可分为两大类：理性主义的基于规则的方法和经验主义的基于语料库的方法。　　本文以军队某部重点科研项目——英汉智能型机器翻译系统为基础，设计实现了机器翻译中的句型转换和译文生成等功能。本文首先论述了课题背景与意义，介绍了机器翻译的发展与研究现状及系统概况。然后对英汉两种语言进行对比研究，论述了英汉语言的特点及差别，并给出相应的消歧策略。接着重点介绍了句型转换和译文生成模块的设计、实现过程。最后给出系统实验结果。　　针对英语中的疑问句等特殊句型，系统采用了利用句型转换对其进行处理的新策略。在格语法的基础上，本文提出了扩展的基于信息的格语法(EICG)，并设计实现了基于EICG的句型转换器，将各种特殊句型转换为陈述句语序。　　翻译是一个高度智能化的过程，单纯的运用某种方法都不能取得比较理想的翻译效果。因此，本文将经验主义的方法和传统的基于规则的方法相结合，在传统的规则体系下，引入翻译模式的支持，两种方法相互补充，设计实现了用于完成源语言的转换和生成工作的译文生成模块。

在基于模式的方法中，基于范例推理的思想，研究了语法信息和语义信息相结合的相似度计算方法。对原有匹配算法进行改进，设计了基于动态规划的句子相似度匹配算法及匹配原则。并给出语义相似度计算公式，通过语义相似度计算来保证对模式进行精确匹配。在基于规则的方法中，针对翻译中遇到的一词多义、介词附着等问题，结合本系统特点，制定了具有本系统特色的翻译规则对各种歧义情况进行处理，完成了短语级目标生成及句子级结构转换等功能。　　在实验阶段，按照国家《机器翻译评测大纲》对系统分别进行了开放性和封闭性测试，由专家对译文质量进行了评估，并对实验结果进行了错误分析。实验表明，系统的译文质量可以达到87.5，翻译正确率可以达到88%。

6.期刊论文 何中军.刘群.林守勋.HE Zhong-jun.LIU Qun.LIN Shou-xun 统计机器翻译中短语切分的新方法 -中文信息学报2007,21(1)

基于短语的统计机器翻译是目前主流的一种统计机器翻译方法,但是目前基于短语的翻译系统都没有对短语切分作专门处理,认为一个句子的所有短语切分都是等概率的.本文提出了一种短语切分方法,将句子的短语切分概率化：首先,识别出汉语语料库中所有出现次数大于2次的词语串,将其作为汉语短语；其次,用最短路径方法进行短语切分,并利用Viterbi算法迭代统计短语的出现频率.在2005年863汉英机器翻译评测测试集上的实验结果(BLEU4)是:0.1764(篇章),0.2231(对话).实验表明,对于长句子(如篇章),短语切分模型的加入有助于提高翻译质量,比原来约提高了0.5个百分点.

7.会议论文 赵红梅.谢军.吕雅娟.刘群 第四届全国机器翻译研讨会(CWMT2008)评测报告(公开版) 2008

为了全面了解国内外机器翻译技术的现状,促进机器翻译技术的研究,根据惯例,第四届全国机器翻译研讨会(CWMT2008)于2008年10月8日到10月22日继续了组织统一的机器翻译评测,以推进参评单位的实质性交流和机器翻译技术的发展。　　本文给出了此次评测的组织、准备过程及结果，为国内外研究单位在机器翻译方面的进一步研究提供了参考数据，本报告内容仅供研究使用，可以在研究论文中引用，但不可用于任何出于商业目的的宣传活动。

8.期刊论文 付雷.刘群.FU Lei.LIU Qun 单纯形算法在统计机器翻译Re-ranking中的应用 -中文信息学报

2007,21(3)

近年来,discriminative re-ranking技术已经被应用到很多自然语言处理相关的分支中,像句法分析,词性标注,机器翻译等,并都取得了比较好的效果,在各自相应的评估标准下都有所提高.本文将以统计机器翻译为例,详细地讲解利用单纯形算法(Simplex Algorithm)对翻译结果进行re-rank的原理和过程,算法的实现和使用方法,以及re-rank实验中特征选择的方法,并给出该算法在NIST-2002(开发集)和NIST-2005(测试集)中英文机器翻译测试集合上的实验结果,在开发集和测试集上,BLEU分值分别获得了1.26%和1.16%的提高.

9.期刊论文 刘洋.刘群.林守勋.LIU Yang.LIU Qun.LIN Shou-Xun 机器翻译评测中的模糊匹配 -中文信息学报

2005,19(3)

目前,大多数机器翻译自动评测方法都没有考虑在未匹配的词语中可能包含被忽略的信息.本文提出一种在参考译文和待评测译文之间自动搜索模糊匹配词对的方法,并给出相似度的计算方法.模糊匹配和计算相似度的整个过程将通过一个例子进行说明.实验表明,我们的方法能够较好地找到被忽略的、有意义的词对.更重要的是,通过引入模糊匹配,BLEU的性能得到显著的提高.模糊匹配可以用来提高其他机器翻译自动评测方法的性能.

10.期刊论文 侯敏.孙建军 汉语中的零形回指及其在汉英机器翻译中的处理对策 -中文信息学报2005,19(1)

回指是语篇衔接的重要手段,零形回指是汉语中常见的一种回指形式.由于汉语、英语是不同类型的语言,因此零形回指对汉英机器翻译会产生一定的影响.本文详细分析了汉语零形回指的确认、类型、产生的原因及使用的条件,指出其对汉英机器翻译造成的主要障碍是生成的英语句子在结构上不合语法,并提出在句组层面上解决问题的算法.