基于 Link Grammar 的英蒙机器翻译系统*

敖其尔[1] 王斯日古楞[2] 吉日木图[1]
（1. 内蒙古大学 内蒙古 呼和浩特 010021； 2. 内蒙古师范大学计算机系 内蒙古 呼和浩特 010022）
E-mail: ochir@imu.edu.cn

摘 要：本文探讨了基于 Link Grammar 的英蒙机器翻译系统的设计与实现算法。文章介绍了英文和蒙文的机器翻译相关基本知识，英蒙机器翻译主要难点和实现方法。最后，举例说明了实现机器翻译的全过程。

关键字：机器翻译；英语；蒙语；Link Grammar

An English-Mongolian Machine Translation System Based on Link Grammar

Ochir[1] Wang-Serguleng[2] Jirumtu[1]
(1.School of Computer Science, Inner Mongolia University, Huhhot 010021, PRC;
2.The Computer Science Department of Inner Mongolia Normal University, Huhhot 010022, PRC)
E-mail:ochir@imu.edu.cn

Abstract: This paper discusses a kind of designing methods and implementing algorithms for the English-Mongolian Machine Translation System . The paper introduced the basic grammar of English and Mongolian related Machine Translation, the main problems of English-Mongolian Machine Translation System and the implementing methods . Finally, gave an example illustrated the whole process of the Machine Translation .

Key words : Machine translation; English; Mongolian; Link Grammar

1 英蒙机器翻译

1. 1 主要内容和研究方法

本文主要研究英蒙机器翻译系统的设计方法和具体实现算法。本系统 EMMT（English-Mongolian Machine Translation）采用了基于规则的方法。在源语言英语的分析中使用美国的 Carnegie Mellon (卡耐基·梅伦)大学计算机系的 John Lafferty 和 *Daniel Sleator* 开发的英语语法分析器—Link Grammar 分析器。该分析器使用连接语法规则和自底

*基金项目：国家自然科学基金项目（60163003）和教育部人文社会科学重大项目（02JAZJD850003）资助。
作者简介：敖其尔（1941~），男，内蒙古大学计算机学院教授，研究方向：蒙文信息处理和计算语言学。

向上的语法分析方法对句子进行分析，把给定的英语句子转化为对应的内部表示形式—语法树。Link Grammar 分析器具有较好的英语语法分析功能，其中 Link Grammar 短语分析器是 Link Grammar 分析器的一部分，它常用的标注与 Penn Treebank 是一致的。我们利用这一点，把一个英语的句子，通过 Link Grammar 短语分析器转换成短语结构语法树。用这棵分析树根据英蒙两种语言的转换规则，把英语语法树转换为对应的蒙古语言的短语结构树，最后根据蒙语的语法规则，形成符合蒙语规则的译文句子。这样可以利用现有的资源，缩短系统开发时间，提高系统分析性能。

1.2 英蒙机器翻译的主要难点

在英蒙机器翻译中我们需要对英语和蒙古语两种语言的语法句法特性进行比较和分析，得出在机器翻译中使用的一种对应关系，从而把一个英语句子转换成对应的蒙语句子。对此，没有任何同样的研究供我们参考。这是一项艰巨的任务，主要体现以下几个方面。

1）英蒙双语词典的建立

众所周知，双语词典的规模直接决定机器翻译系统的性能。词典的规模越大，包含的单词范围越广，系统的性能就越好。然而，自然语言是个无限集，我们不可能建立一个包含所有词的词典。加上人力和资金的限制，我们选择了蒙古族中学的英语教材和英汉蒙学生使用词典以及参考其它有关词典，建立了英蒙双语词典，对每个词条标注了相关的属性。

2）英文到蒙文的转换规则的设计

我们使用了基于规则的转换方法，因此，研究从英语到蒙语的转换规则是我们整个翻译系统的核心。要正确的建立规则库，必须对英语和蒙语两种语言的语法、句法规则作深入的分析，从中抽取有用的知识，用形式化的方法表示它。

3）英语时态与蒙语动词的对应关系

英语时态是英语的一个重要语法现象，理解一个句子时 ，我们必须考虑它的时态信息。然而，蒙语的时态是通过对句子中动词的词干加上不同的后缀来体现的。那么，如何将英语的 16 中时态转换成蒙语的正确动词形式也是我们要解决的一个难题。

4）英语介词与蒙语的关系

在英语的短语结构中，介词短语占着很大的比例。在 EMMT 系统中，把英语的介词翻译成蒙语时，往往与蒙语的“后置词”、“助动词”和“格的附加成分”等相对应。因此，怎样正确全面的处理这类问题，也是本文要解决的一个难点之一。

5）英蒙转换算法和蒙语生成算法

转换生成算法是实现系统的关键。如何在转换阶段充分利用 Link Grammar 分析器的分析过程中所获得的有用信息，并且把分析过程和转换过程有机的连接起来是我们在转换算法中必须考虑的问题。同时，蒙古语言是一种比较特殊和复杂的语种，如何生成符合蒙语语法句法规则的句子，与我们编写的转换生成算法直接相关。

正确有效的解决以上问题，是 EMMT 系统的关键。这些问题的解决直接依赖于对两种语言的对比和研究及对 MT 方法的深入分析。

2 英语和蒙语的语法概要

2.1 英语语法概要

在 EMMT 系统中，我们必须对作为源语言英语进行深入的分析，才能够得出它的正确的语

法树，同时充分利用分析过程中获得的句子语法，才能生成正确的蒙语译文。下面就 EMMT 系统中涉及到的主要英语语法作简单的说明。这里所介绍的英语语法是面向自然语言处理和 MT 研究的英语语法。

1）英语词性分类

词是构成句子的基本单位，在英语中，根据形式、意义及其在句子中的功能通常分为十大词类。其名称及缩写如下：名词：n　代词：r　形容词：adj　数词：num　动词：v　副词：adv　冠词：art　介词：prep　连词：conj　感叹词：int

2）英语短语

短语是由词构成的。短语是意义上自成一个单位，但不构成句子或从句的语法。短语的种类很多，英语中常用的有五类，其名称及缩写如下：名词短语：np　动词短语：vp　形容词短语：ap　介词短语：pp　副词短语：dp

3）英语句子结构

句子是词的序列，词构成短语，短语构成句子。英语的句子成分主要有：主语（subject），谓语（predicate），宾语（object），定语（attribute），状语（adverbial adjunct）和补语（complements）。把定语和状语称为修饰成分（modification）。

通常情况下，主语由名词短语或从句构成，谓语由动词短语构成，补语由名词短语或形容词短语或从句构成，修饰成分一般由形容词短语、副词短语、介词短语和从句构成。一个英语的句子结构可用一个括号序列表示出来。如：He saw a big black dog in the park. 的结构如下：

```
(s (np He)
   (vp saw
       (np a big black dog)
       (pp in
           (np the park)))
   .)
```

4）英语基本句型

由主语、谓语动词、表语、宾语、宾语补助语等组成的英语句子，以其组合方式可分为如下 5 种基本句型：

（1）　主语+不及物动词

（2）　主语+及物动词+宾语

（3）　主语+系动词+表语

（4）　主语+及物动词+间接宾语+直接宾语

（5）　主语+及物动词+宾语+宾语补助语

2.2 蒙语语法概要

在 EMMT 系统中，蒙古语是目标语言。因此，我们只简单介绍英语到蒙语的转换和蒙古语生成有关的语法知识。蒙古语言文字具有悠久的历史。蒙古语是很早以前居住在我国北部和中亚地区的蒙古部落使用的语言，属于阿尔泰语系蒙古语族语言：是从上到下由左向右竖写的拼音文字。

1）蒙语词性分类

根据词的意义把蒙语的词分为静词类、动词类和无变化词类等三大类。其中名词类词指示事物、时间、地点、性质和数量等，具有数、格和领属的变化。静词类中有：名词、形容词、数量词、代词、时位词等五个小类。动词类按其词汇意义首先可分为两类：即实

义动词与虚义动词。实义动词具有具体的词汇意义，句中表示主要的动作和行为，普通的动词都属于实义动词。虚义动词是辅助性的动词，主要表示抽象的语法意义，数量很少。实义动词是从过程方面表示行为状态的普通动词。实义动词除了原有的动词外，还可以从别的词类派生大量的派生词。所以动词的种类及数量繁多。蒙语动词类的形态变化也比较复杂，它除了具有词尾变化以外，还有词干中的形态变化。

2） 蒙语短语

蒙古语的短语可以分为：名词短语：NP，动词短语 VP，形容词短语：AP，数词短语：MP，副词短语：DP

3） 蒙语句子结构

句子一般由主语部分、谓语部分组成。主语和谓语是句子的主要成分，次要部分包括定语、状语和宾语。在蒙古语中宾语一般在谓语之前。

3 英蒙机器翻译系统的设计

3.1 设计方法

在 EMMT 系统中我们使用了基于规则的方法。

源语言英语的分析用美国的 Carnegie Mellon (卡耐基·梅伦)大学计算机系的 John Lafferty 和 Daniel Sleator 开发的英语句法分析器—Link Grammar。该分析器使用连接语法规则和自底向上的语法分析方法对句子进行分析，把给定的英语句子转化为对应的内部表示形式—语法树。Link Grammar 分析器具有较好的英语语法分析功能，其中 Link Grammar 短语分析器作为 Link Grammar 句法分析器的一部分，它的常规用的成分与 Penn Treebank 是一致的。我们利用这一点，把一个英语的句子，通过 Link Grammar 短语分析器转换成短语结构树。再把这棵分析树通过英蒙两种语言的句型转换规则，转换为对应的蒙古语言的短语结构树，最后根据蒙语的生成规则，形成符合蒙语规则的句子。

3.2 EMMT 系统的设计

翻译系统采用了基于规则方法，系统包括分析、转换和生成三部分。 翻译系统的流程如图 2。

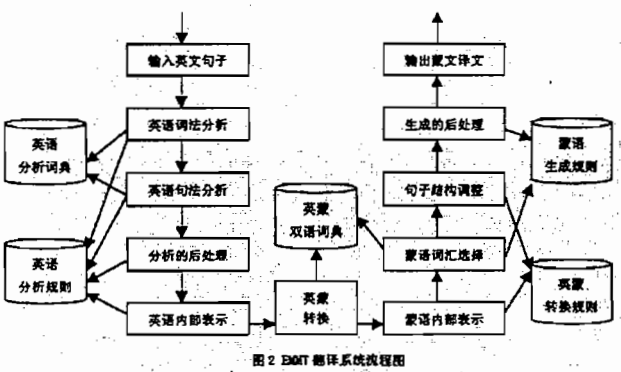


图 2 EMMT 翻译系统流程图

在分析、转换和生成的各个阶段，都要有相应的知识库的支持。实现 EMMT 系统，首先

要建立一个正确可靠的知识库，知识库包括词典库、规则库和语料库。

4 英蒙机器翻译系统的实验

EMMT 系统是使用 C++ Builder 5.0 开发的。双语词典和转换生成规则以数据库形式存储。双语词典包含了九年义务教育三年制初级中学教科书中出现的所有单词。通过 Link Grammar 短语分析器对英语句子分析的结果和语法分析树的生成，可以获得给定句子中使用的英语分析树。按照转换生成语法的原则，通过反复观察和试验，形成英蒙转换生成规则，再利用所获得的规则对统一类型的其他句子进行处理。

4.1 具体试验的过程

EMMT 系统把一个输入的英文句子，通过分析、转换和生成三个阶段后，可以获得对应的蒙文译文。以下就是通过几个简单的例子来展示一下 EMMT 系统翻译试验过程。

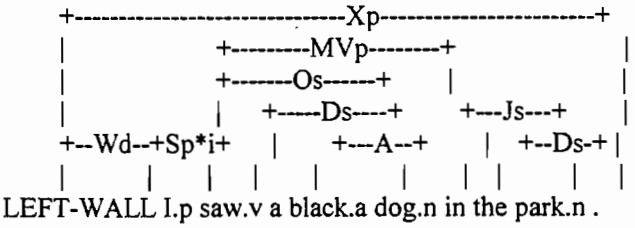
1) 一般陈述句

例如，我们处理英语句型：主语+及物动词+宾语：S+V+O。

输入句子： I saw a black dog in the park.

分析阶段：

获得如下 Link Grammar 语法树：



输出对应的短语结构语法树为：

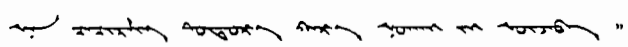
```
(S (NP I)
  (VP saw
      (NP a black dog)
      (PP in
          (NP the park)))
  .)
```

转换阶段：

输出所用的转换规则：

```
s->np vp w=>S(NP/np VP/vp W/w)
np->r=>NP(R/r)
vp->v np pp=>VP(PP/pp NP/np CF V/v)
pp->p np=>PP(NP/np O/p)
np->t n=>NP(M/t N/n)
np->t a n=>NP(A/a N/n)
```

生成阶段：

输出译文： 

2) 过去完成时句子

输入句子：I had studied in the classroom.
分析阶段：
获得如下 Link Grammar 语法树：

```
     +------------------------Xp-----------------------+
     |                                  +----Js----+    |
     +--Wd--+Sp*i+--PP--+-MVp-+   +---Ds--+   |
     |      |     |         |      | |         |    |
LEFT-WALL I.p had.v studied.v in the classroom.n .
```

输出对应的短语结构语法树为：

```
(S (NP I)
   (VP had
      (VP studied
         (PP in
            (NP the classroom))))
   .)
```

转换阶段：
输出所用的分析和转换规则：
s->np hp w=>S(NP/np VP/hp W/w)
np->r=>NP(R/r)
hp->v vp=>VP(VP/vp)
vp->v pp=>VP(PP/pp V/v)
pp->p np=>PP(NP/np O/p)
np->t n=>NP(M/t N/n)
生成阶段：
输出译文：

4.2 结 论

我们在研制 EMMT 系统时，编写了包含有八千多条关键词的英蒙双语机器翻译词典，建立了含有两百多条转换规则的规则库，编写和调试了有关英蒙转换和蒙文生成的程序，为蒙文机器翻译的研究打下了基础。实际上，这些工作对于汉蒙机器翻译、俄蒙机器翻译和日蒙机器翻译当中均可用到。

英蒙机器翻译的研究是一项基础性研究工作。它对少数民族地区的科学技术，文化事业和经济发展具有重要的实际意义，对蒙古语言文化走入世界舞台打下了必要的基础。今后，我们打算在此基础上，继续完善系统，扩充词典、语料、和规则等资源，使得本系统早日走出实验室。

参考文献

[1] 赵铁军，机器翻译原理，哈尔滨工业大学出版社，2000
[2] 姚天顺，自然语言理解，清华大学出版社，1998
[3] 清格尔泰，现代蒙古语语法，内蒙古人民出版社，1992
[4] 冯志伟，自然语言机器翻译新论，语文出版社，1994
[5] 敖其尔，从英文到蒙文的机器翻译，内蒙古大学学报(哲学版)，1988，3：39-50
[6] http://www.link.cs.cmu.edu/link.