英语句法分析树向汉语分析树的转换

姚建民 张晶 赵铁军 于浩 李生

哈尔滨工业大学计算机科学与技术学院，哈尔滨 150001

E-mail: james(zhj,tjzhao,yh)@mtlab.hit.edu.cn

摘要： 介绍了 MTS2000 英汉机器翻译系统中译文选择和转换生成的实现策略。基于句法结构的译文选择将句法上下文集合视为词包，统计集合中各词的词信息和词性信息作为上下文特征，以 Bayes 最小错误概率公式作为评价函数选择译文。基于规则的转换生成模块在句法分析树的基础上，利用词性、语义，甚至词特征进行源语言句法树向目标语句法树的转换。

关键词 译文选择 转换生成 机器翻译

English Chinese Transfer of a Syntactic Parsing Tree

Yao Jianmin, Zhang Jing, Zhao Tiejun, Yu Hao, Li Sheng

School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001

E-mail: james(zhj,tjzhao,yh)@mtlab.hit.edu.cn

Abstract: The translation choice and transfer modules in the English Chinese machine translation system of MTS2000 are introduced. The translation choice is realized on the basis of a grammar tree and takes the context as a word bag, with the lexicon and POS tag information as context features. The Bayes minimal error probability is taken as the evaluation function of the candidate translation. The rule-based transfer and generation module takes the parsing tree as input and operates on the information of POS tag, semantics or even the lexicon.

Keywords Translation choice, Transfer and Generation, Machine Translation

1. 引言

当今社会处于信息时代，Internet 网的迅猛发展，迫切需要通过机器翻译消除不同国籍人们之间的文字障碍。但是，在人们对机器翻译充满希望的同时，还必须认识到，自然语言翻译是人类高级智能活动之一，而人工智能的研究尚未达到完全理解自然语言的水平。一般来说，机器翻译包括三个子系统[1]：（1）分析：对输入的源语言进行分析，形成带有句法功能标记的层次性的句法分析树；（2）转换：把源语言句法分析树映射到生成目标语的生成树；（3）生成：根据生成树生成目标语；

哈工大机器翻译实验室开发的 MTS2000 系统是统计和规则相结合的双向机器翻译系统。图 1 为该系统的处理流程。MTS2000 采取了分析和转换分开的策略，模块化的构造有利于统计方法和规则方法的有机结合，并为融合新的技术提供了方便的接口。转换生成包

含了句法分析之后的两个模块。具体任务是：给定一棵英语的句法树（见图 2），以语言的语法特征、简单的语义特征为知识源，生成一棵汉语的句法树，汉语树的所有终结点按顺序构成汉语译文的词序列。

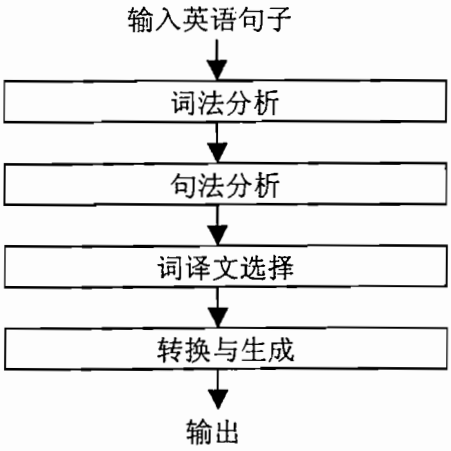

图 1　英汉机译系统流程

输入的源语言句子经过词法分析，词性标注和句法分析器，形成一棵具有层次性，带有句法功能标记的句法分析树[2]。例如：

输入英语句子“Sell one’s interest in the company.”

句法分析之后，我们将得到树状结构如图 2，等同于如下形式：（S(VB(Sell/VB NP (one’s/PRP$ interest/NN)) PP(in/IN NP(a/ART company/NN)).))。

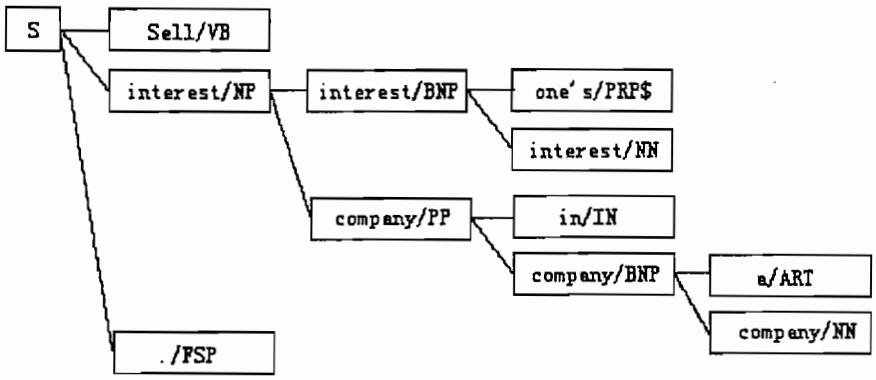

图 2.　句法分析器输出示例

目前，我们的英语句法分析器已经能够生成比较完整的句法分析树。而英语句法分析树作为转换生成的输入，包含了源语言句子中各词节点间相互依赖关系的基本信息，也包括各词节点的部分语义信息。这些信息是转换生成的起点。

句法分析后，转换生成的任务包含了歧义词的译文选择、词序调整及部分功能词的增删。转换与生成部分被分为两个模块：其一用于译文选择，其二用于结构转换和汉语生成及译文的修正。

2. 基于句法分析的译文选择

首先，我们对机器翻译中歧义词的译文选择做形式化描述[3]：假设待翻译的源语言句子 ES，其中歧义词 EW 有 M 种目标语译文 CW1，CW2，... CWM，在一定上下文特征 C 的条件下，出现 CW1，CW2，... CWM 的概率分别为 P（CW1 | C）、P（CW2 | C），...，P（CWM|C）。根据 Bayes 最小错误概率公式：

CW = argmax[P(CWk |C)] = argmax[logP(CWk) + logP(C|CWk)]　　（1）

当不失一般性地假设 P（CW1 | C）> P（CW2 | C）>，...，> P（CWM|C）成立时，我们选择译文 CW1 作为 EW 的译文。根据 Naïve Bayes 假设公式：

P(C|CWk) = P({vj | vj in C}|CWk) = ΠVj in C P(vj|sk)　　（2）

公式（1）改写为：

CW = argmax[P(CWk |C)]= argmax[logP(CWk) + ΣVj in C logP(vj| CWk)]　　（3）

其中，P(CWk) 表示译文 CWk 在语料库中的出现概率；P (vj| CWk)表示上下文特征 vj 与译文 CWk 同现的概率。

从上面的形式化描述看出：统计方法选择译文的关键是寻找恰当的上下文，选择上下文特征 vj。已有研究常用的方法是定义一定大小的上下文词窗口（word window），即认为在歧义词某个窗口范围内的单词对词义的选择有贡献。例如李涓子采用长度为 6 的词窗口用于词义排歧[4]；荀恩东[5]定义长度为 4 的移动窗口；Hwee Tou Ng 采用与歧义词偏移±2 的词窗口[6]。但是这种将上下文定义在歧义词一定距离以内的方法可能导致两个问题：（1）对词义选择有贡献的上下文没有被词窗口所覆盖；（2）位于词窗口内部的词汇对词义的确定并无贡献，并且带来噪声。通过对语料库中大量的歧义词进行考查，我们选择上下文时考虑了句中单词与歧义词在句法结构上的相关性，而不是二者之间的距离。具体阐述如下：

从以上的分析，我们利用了基于句法结构的译文选择方法：将译文选择置于机译系统的句法分析器和转换生成器之间，获取歧义词的句法上下文集合。选择译文时，将句法上下文集合视为词包，即不考虑单个词对译文的贡献，统计集合中各词的词信息和词性信息作为上下文特征，以 Bayes 最小错误概率公式作为评价函数选择译文。

本文使用句法上下文，统计与歧义词在结构上有关联的特征，有效地利用了句法分析的结果，特点在于（1）无需人为定义上下文窗口的大小，而且将获得尽可能多的对词义选择有贡献的上下文特征；（2）滤除与歧义词在句法结构上无关的上下文特征；（3）这些特征考虑的是与歧义词在结构上的关联，并不要求完全正确的句法分析结果。从上述特征，我们看出这种方法使用上的可行性。

3. 基于规则的转换和生成及译文的修正

机器翻译的根本任务是实现一种语言到另一种语言的意义等价的转换。它并非如自然语言理解那样只是对一种语言的操作。作为一个机器翻译系统，它既要考虑到源语言的语法和语义规律，还要考虑到目标语言的规律，忽略了任何一方都是不行的。换句话说，双语互译规律的发现和应用才是机器翻译的最本质特征。目前的机器翻译系统基本有三种方法[7]：1）基于模式/规则的系统：由产生式规则构成系统知识库的主体。规则或模板一般由人工编制或从语料中获得；2）基于实例的机器翻译方法。给定一个源语言的输入片段 S（短语或单词）和一个双语文本集，其中包括源语言的片段文本 S’和对应的译文片段 T’，将 S 与双语文本集中的源语言端的片段进行匹配。选出“最相似”的片段，其相应的译文片段或其修正作为 S 的翻译。3）基于统计的机器翻译方法：是一种基于单语语言模型和双语对齐模型的方法，通过统计大规模的（双语对齐）语料库可以得到这些概率。

对 MTS2000 而言，结构转换是从英文的句法分析树出发，构造汉语句子的树结构，而汉语生成即从汉语树结构生成汉语词链，构成译文句子[8]。该模块采用了基于规则的知识表示。规则系统设计的好坏直接关系到机译系统性能。

英汉机译系统的规则描述语言是以产生式的形式描述的，即分成条件部分和操作部分。条件部分作为可变长的扫描窗口，给出短语或某种语言特征的上下文约束条件，操作部分生成相应的译文或给出条件部分的生成特征。如果条件满足，则按操作部分进行操作。规则的这种表达方法体现了系统的一个技术特点，即转换与生成一体化。规则描述语言力求接近自然语言，尽量做到与人的逻辑习惯相一致，并且可用多种手法，多种方式描述。

规则的条件部分就是由多个节点号连同它的节点规则项用“＋”符合连结而成的符号串。规则的操作部分由对应条件部分的各个节点规则项确定的译文和动作函数组成。

例如，形容词和一个名词合并为一个名词短语的规则为：

```
0: Cate=A + 1: Cate=N ->
0: * + 1: * + _NodeUf(N, 0, 1)
```

其中“*”表示相应节点译文，_NodeUf 为节点合并函数，在节点合并的同时生成了新的译文。

概括起来，英汉机器翻译系统的转换生成规则表示具有如下的特征：

①描述语法语义特征的是一个由框架名和框架值构成的字串，中间由等号连接.

②条件之间可进行与、或、非的运算。

③可以对句子同层节点前后任意扫描和测试。

④动作函数和测试函数可以生成传递特征或测试相应特征。

系统的规则描述语言基本上可以描述大量语言现象，并且做到了简明易懂，对系统起到了重要的支持作用。

语言规则是有层次的。规则放在以词性类为特征入口的规则库中。规则库中的规则是有优先级的，这个优先级决定了规则匹配的先后次序。一般来说，越是具体的规则，其优先级越高；越是抽象的规则，其优先级越低。因此，在规则库的相同入口下，我们把更长

的、约束更多的、更具体的规则放在前面。规则的层次性有利于解决规则冲突。
模块名：基于规则的英汉转换和汉语生成
把转换生成规则读进内存，构造成匹配树形式；
读入一句英语；
如果句子的部分句法树满足规则的条件部分，根据规则的执行部分处理该句法树；
循环，直到无规则可匹配；
句子转换和生成规则的功能如下：
词义选择，即为歧义词选择恰当的译文，例如

```
0:C=CD + 1:Sem=Year + 2:W=old ->
 0:* + I:岁+ _NodeUf(N,0,1) + _PutFeat(N,H=PP)@
0:C=ART + 1:W=Chinese + 2:H=H->
0:一个 + 1:中国的 + 2:*@
```

英语句法树转换生成汉语句法树，即调整词序和词间依赖关系，例如：

```
0:C=WP + 1:W=be|W='s|W='re + 2:C=N|C=DT|C=PRP->
2:*+1:是+0:*@
```

下面是转换生成规则中我们利用的词的特征和定义的操作符：
1. 词性，e.g. C=N, 表示词性为名词
2. 词，e.g. W=go, 表示词为 go, goes, going, went, gone, 等
3. 语义特征。这些特征可以从词典中获取（包含语义的大类、中类和小类），也可以在规则库中由规则生成（可以生成表示句法特征或语义特征，或二者综合）。
例如，0:W=of+1:C=N->0:+1:*+_NodeUf(1,0,1)+_PutFeat(N,H=OF)@
表示：（0:W=of ）of （+1:C=N）加名词（或短语）（->0:+1:*）则译文为名词的译文，（+_NodeUf(1,0,1)）并且两个节点合并，核心为后者，（+_PutFeat(N,H=OF)），合并后的新节点赋予特征 H=OF.
又如，0:W=solution+1:H=OF->1:*+I:的+0:解答+_NodeUf(0,0,1)@
表示：词 solution 后接有 H=OF 特征的节点，则译文为：“解答”。
操作符说明：
1. &:与关系
例如，0:W=have+1:Head=Be+2:C=V&CX=ing->I:一直在+2:*+_NodeUf(2,0,1,2)@
节点 2 的词性上动词， 并且有特征 CX=ing……
2. |:或关系
例如，1:C=JJ+2:W=desk|W=table|W=face->0:*+I:张+1:*+2:*@
表示：节点 2 的词为 desk 或 table 或 face……
3. *:默认译文
例 如 ， 0:W=January|W=February|W=March|W=April|W=May|W=June|W=July|W=August|W=September|W=October|W=November|W=December+1:C=CD->0:*+1:*+I: 日 + _NodeUf(0,0,1) +_PutFeat(N,H=DATE)@
表示取词典中的第一译文项。
4. I:插入译文

如上例，+I:日+ _NodeUf(0,0,1)，表示插入“日”。

函数说明：

1. _PutFeat()

例如，0：C=CD+1:W=o'clock->0:*+1:点钟+_NodeUf(N,0,1)+_PutFeat(N,H=M)@

表示，合并两个节点，并赋予新节点特征 H=M。

其他函数：

2. _NodeUf()

合并节点，0：C=CD +1:Sem=Year+2:W=old->0:*+I:岁+_NodeUf(N,0,1)+_PutFeat(N,H=PP)@

表示将前两个节点合并，并合并为新节点，新节点特征 H=PP。其它次要函数从略。

4. 总结和讨论

我们的系统原型已经完成，大规模的开发正在进行。从目前的工作看来，知识的获取和知识库的组织会成为 MTS2000 这样面向非受限域自然语言翻译的瓶颈。译文选择阶段的知识获取要求大规模词对齐的双语语料库，并对语料库中歧义词的数量有依赖性。我们正在与微软中国研究院联合开发的对齐双语语料库将对本模块的质量产生提高。转换规则库的开发随着知识库的扩大产生更大的知识冲突和部分冗余。研究规则库自组织技术也会对转换生成的质量产生推动作用。

参考文献：

[1] 赵铁军 荀恩东 陈斌 刘小虎 李生，基于目标语统计的译文选择研究，应用基础与工程科学学报，1999，7 (1)：101-110

[2] Meng Yao, Zhao Tiejun, Yu Hao, Li Sheng, A Decision Tree Based Corpus Approach to English Base Noun Phrase Identification, Proceedings International conference on East-Asian Language Processing and Internet Information Technology, Shenyang, 2000: 5-10

[3] Christopher D. Manning, Hinrich Schütze, Foundation of Statistical Natural Language Processing. The MIT Press. pp229-262. 1999.

[4] 黄昌宁 李涓子. 词义排歧的一种语言模型. 中国语言学会第十届年会论文.1997.10 福州

[5] 荀恩东，李生，赵铁军. 基于汉语二元同现的统计词义消歧方法研究，高技术通讯. 1998，10(8)：21-25

[6] Hwee Tou Ng. Exemplar-Based Word Sense Disambiguation: Some Recent Improvements. In Proceedings of the Second Conference on Empirical Methods in Natural Language Processing (EMNLP-2), August 1997

[7] 赵铁军等，机器翻译原理，哈尔滨工业大学出版社，2000

[8] 姚建民，张晶，转换和生成模块的设计与实现，哈尔滨工业大学机器翻译研究室技术报告，2000