一个汉英机器翻译系统的计算模型与语言模型*

刘 群+ 詹卫东++ 常宝宝++ 刘颖+
(+中国科学院计算技术研究所二室 北京100080)
(++北京大学计算语言学研究所 北京100871)

摘要：本文介绍我们所设计并实现的一个汉英机器翻译系统。在概要介绍本系统的主要目标和设计原则的基础上，着重说明系统的计算模型和语言模型，最后给出实验结果和进一步的打算。

关键词：自然语言处理 机器翻译 中文信息处理

一、引言

我国的机器翻译研究近年来取得了很大的发展。特别是英汉机器翻译系统的研制已经取得了较大的成功，达到了初步实用的阶段。相对而言，汉英机器翻译的研究却进展比较缓慢，离实用化还有相当的距离[1]。我们的目的是利用目前最新的计算机软件技术、相对成熟的机器翻译方法和先进的汉语语法理论，构造一个初步实用的汉英机器翻译系统。本文将对我们所开发的系统所采用的计算模型和语言模型作一个总体性的介绍，而不涉及过多的细节。

下面我们简要介绍一下本系统的几个主要设计原则：

⑴ 采用成熟的技术

我们的目的是构造一个真正实用的汉英机器翻译系统，因而在可供选择的若干技术路线面前，我们将尽量选用比较成熟的技术，而在现有技术难以解决问题时再尝试一些新技术。

⑵ 开放的体系结构

开放的体系结构主要体现在系统的实现上所采用的软件构件技术[8]。整个系统采用一些相对独立的软件构件组成，因而可以方便地对系统进行修改、维护和扩充。翻译的过程严格按照独立分析、独立生成的原则进行组织，每一阶段的算法相互独立，对其中一个阶段算法的修改不会对其他算法造成影响。

⑶ 方便的调试环境

本系统强调为语言工作者提供一个方便的调试环境。系统提供多窗口图形界面的知识库调试工具，支持课题组中多人同时通过网络对一个知识库进行操作。提供对翻译过程直观显示，用户可以清晰地看到翻译过程的每一步操作。提供翻译出错原因查找机制，用户可以轻松确定翻译出错的位置。

机器翻译系统可依据不同的标准进行分类，这些标准也刻划出本系统的一些基本特点：

(1) 规则方法与语料库方法

规则方法发展到今天，相对来说已比较成熟，但由于专家描述的规则知识通常颗粒度较大，不利于处理大量的细节，因而在处理大规模的开放语料时，遇到了难以克服的困难；而从预料库中获取的知识颗粒度较小，在自然语言处理的某些方面取得了成功，但纯粹基于语料库的的机器翻译系统，还没有比较成功的例子。本系统目前采用的是基于规则的技术，我们计划将其扩展成为一个规则方法与语料库方法相结合的系统。

(2) 转换方法与中间语言方法

从理论上说，在实现多种语言互译的机器翻译系统时，中间语言方法可以节省很多的工作量。但从已实现的系统来看，使用转换方法较易取得成功。本系统也采用转换方法。

(3) 确定性算法与不确定性算法

确定性算法的优点是算法较为简单，翻译速度快，缺点是不能提供回溯的能力，翻译过程任何一步的错误将导致整个翻译的失败。不确定算法刚好相反。本系统采用不确定性算法，翻译过程的每一步骤都是不确定的，都可以回溯。

二、计算模型

我们从系统结构、知识表示、翻译算法三方面来介绍我们所采用的计算模型。

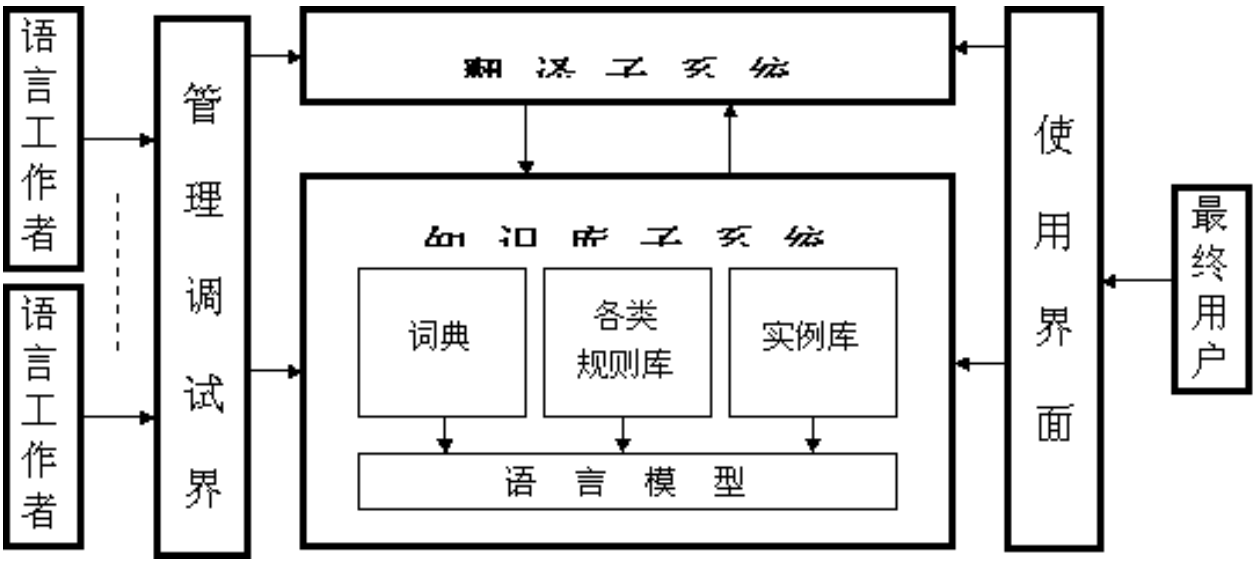

图1 汉英机器翻译系统总体结构

1、系统结构

系统总体结构如图1所示。

2、知识表示

机器翻译的过程可以看成是一个运用知识进行推理的过程。知识表示是这一过程的基础。我们把机器翻译中用到的知识表示形式分为内部知识和外部知识两类。其中外部知识是存放于知识库之中，由语言工作者进行管理的知识，

如词典和各类规则库等，内部知识是翻译过程中临时生成的，用于描述所翻译的句子的语法语义特征的知识，如树形图、特征结构和语义网络等。

本系统的外部知识表示由知识库子系统进行处理。知识库包括一个语言模型、一部词典、多个规则库和一个实例库。

本系统设计的严格的语言模型起统帅作用，其中规定了本系统所使用的源语言和目标语言的词法模型、句法模型和语义模型，即词法、句法和语义的分类和各种属性描述。所有知识库中所用到的各种语言知识描述用的符号格式都必须符合语言模型中的规定。

整个系统使用一部双语词典。

多个规则库对应于翻译的各个步骤，每个步骤使用相应的规则库。每个规则库的具体格式各不相同，但基本上都采用“树结构+约束”的形式。在知识库的格式定义上，我们特别强调不仅要能描述全局性知识，也要能描述一些局部性的知识。因此我们特别强调词典的描述能力。例如，词典中的局部规则与全局规则具有完全相同的格式，在使用上局部规则优先于全局规则，这样特别有利处理一些与具体词相关的特殊用法。

实例库用于存放系统翻译过的句子及其相关信息。

本系统的内部知识表示形式包括线图(Chart)、树结构和特征网络三种形式。

线图源于Chart Parsing算法，是一种比较通用的语言内部结构表示方法，可以同时表示翻译过程中产生的大量词结点和短语结点，也可以适应多种不同的分析算法。

树结构是短语结构分析中最常用的一种表示方法，用于描述句法成分（包括词结点和短语结点）之间的组合关系。每个树结点对应于线图中的一个词结点或短语结点。我们所使用的树结构表示法中要求标出每个句法成分的中心子结点，用于处理属性值在句法成分之间的传播。

特征网络是本系统所使用的一种特有的知识表示方法。这种表示法融合了特征结构表示法[4]和语义网络表示法的一些特点并加以改进，以适合汉英机器翻译的需要。具体来说，特征网络表达具有以下特点:

1、一个特征网络由许多个互相关联的特征结点所组成;

2、一个特征结点是若干个特征的集合，一个特征是一个“属性-值”对;

3、属性分为简单、原子属性和关联属性两种，原子属性的值是一个原子，关联属性的值是另一个特征结点;

4、原子分为层次型、符号型、数值型、布尔型等多种类型，原子之间可以通过与、或、非等逻辑操作构成复杂原子，每一种类型的原子有不同的合一算法;

5、特征结点之间通过关联属性互相连接，这种连接可以构成回路，我们改进了合一算法，使得这种回路不至于在合一运算时造成死循环;

6、一个特征结点对应着句法分析中已经出现或可能出现的一个句法成分，而每一个句法成分（即句法树中的结点）一定有唯一的一个特征结点与之对应;

7、在一定的条件下，属性的值可以在特征结点之间进行传递;

8、特征结点之间实行真正的合一运算，而不是伪合一运算。

特征网络表示法作为一种最基本的知识表示方法在本系统中发挥着重要的作用，它基本上满足了我们在汉英机器翻译中描述各种复杂的语言现象的需要。

3、翻译算法

我们采用基于转换的翻译方法，遵循独立分析、独立生成的设计原则[3]。

具体的翻译流程下图2所示。

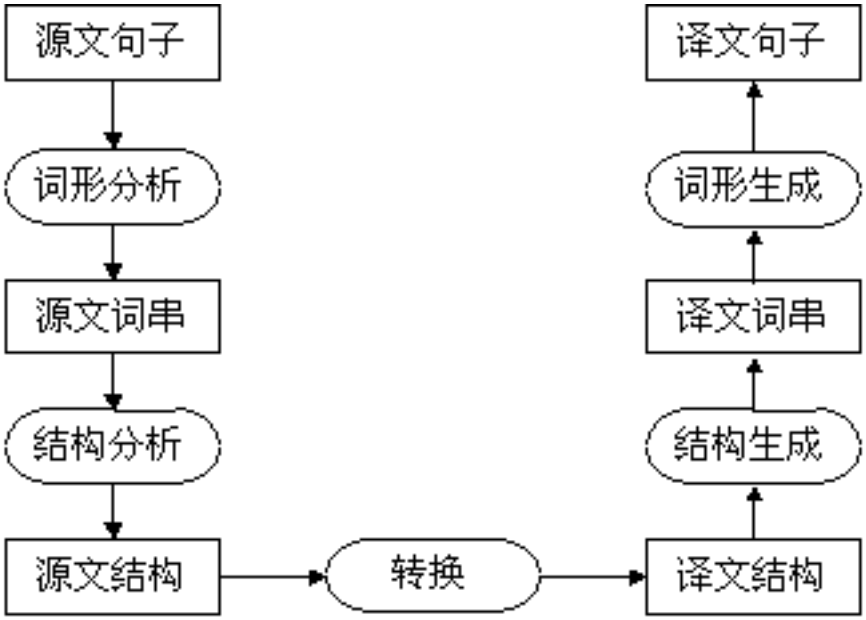

图2 翻译的流程

其中，汉语的词形分析阶段分为重叠词处理和切分两个步骤，汉语的切分采用双向最大匹配算法。出现切分歧义时，不做判断，保留到结构分析阶段进行处理。结构分析阶段采用改进的Chart Parsing算法[6]。转换阶段采用自顶向下与自底向上相结合的局部子树变换算法。结构生成阶段采用自底向上的局部子树变换算法和自顶向下的全局子树位移算法。

4、 用户界面

用户界面包括使用界面和管理调试界面。由于本系统还处于开发阶段，我们着重开发了管理调试界面。对于一个实际的机器翻译系统来说，语言规则和词典的调试工作是非常重要的。一个良好的机器翻译系统管理调试界面，可以使语言工作者直观方便地进行语言规则和词典的调试，大大提高调试的效率，进而有效地提高语言知识库的质量。

本系统的管理调试界面分为知识库管理界面和翻译调试界面两部分。知识库管理界面帮助语言工作者对各知识库进行直观的创建、管理和查询等操作，翻译调试界面帮助语言工作者通过观察具体例句的翻译过程对语言知识库进行调试。系统依次以可视的图形显示源文正文、分析产生的每一个源文词语结点、源文短语结点、源文句法树、转换生成产生的译文句法树、译文词结点和最后的译文正文。语言工作者可以根据需要显示任何一个句法成分的产生过程及其对应的特征结点（属性和值）。

三、语言模型

在一个机器翻译系统中，计算模型决定了该系统的能力的极限，即该系统最好能翻译到什么程度；而语言模型则决定了该决定了这种极限能在多大程度上实现。没有好的语言模型，系统的计算模型再好，也不可能得到高质量的译文。

本系统采用以语法分析为主，以语义分析为辅的语言模型。

就汉英机器翻译系统而言，目前还没有专门适用于机器翻译的汉语语法模型。本系统采用的语言模型主要来源于北京大学计算语言学研究所研制的《现代汉语语法信息词典》[2]（以下简称《词典》），并在该词典所采用的语言模型基础上修改扩充而成。

1、汉语词语分类和属性[10]

本系统所采用的汉语词语分类和属性取自于《词典》，并作了少量的改动。《词典》中将现代汉语词语（包括标点符号、语素、成语等）分为26类，我们只采用了其中的20类，并将其余6类归并到这些类中。《词典》中有大量的属性描述，我们根据机器翻译的需要对这些属性作了一定的取舍，并增加了少量新属性。本系统所使用的机器翻译词典就是在《词典》的基础上修改扩充而成。在使用中我们体会到，《词典》对现代汉语词语的分类合理，对词语语法功能的描述非常详尽，基本上能满足汉英机器翻译的需要。

2、汉语短语分类和属性[7,10]

对汉语短语的分类，我们继承了《词典》中对汉语词语分类时采用的“功能分类”思想，将短语（包括句子）分成np,vp,ap,tp,sp,dp,pp,mp,mcp,dj,fj,zj等12类。另外，我们还定义了内部结构、语气、被动、否定等短语属性。

我们认为，短语和汉语词语一样，采用按功能分类的思想，而不是按结构分类或按功能-结构混合分类，是符合机器翻译用汉语语法体系要求的。这是因为，功能反映了一种短语与其它短语互相结合的能力，而语法规则所描述的就是短语之间如何互相组合构成新的短语，因而采用功能分类是非常自然而贴切的。短语的结构从本质上说只是短语内部成分之间的组成关系，虽然结构对功能也有一定的影响，但它并不直接反映短语向外结合的能力。因此我们只是把短语的内部结构作为一种属性来对规则进行约束，而不是作为分类的依据。在实践中我们感觉到这种做法是恰当的，既不至于导致规则的描述能力不够，也不会产生大量的冗余规则。

总的来说，我们对汉语短语的认识要比我们对汉语词语的认识肤浅得多。在很多情况下，我们没有足够准确的属性来描述规则的约束条件，尤其是一些很常用的歧义结构，如np+np,vp+vp,np+vp等等。这尤其需要我们机器翻译研究工作者与语言学家共同努力，对汉语短语的语法功能进行更加深入的研究。

3、语义分类和属性[9]

本系统是一个以语法分析为主，语义分析为辅的系统。虽然如此，在本系统中，为消解句法分析和转换时的歧义，语义分析还是起着重要的作用。

本系统采用的语义模型主要包括语义分类和配价分析[5]两个方面。

我们制定了一个比较详尽的语义分类体系，对每一个汉语实词都要填写其相应的语义分类，而对于名词、动词、形容词三类词语还要填写配价数以及相应配价成分的语义类。在规则的约束条件中，对某些短语的组合规定了一定的配价关系，如果这种关系不能被满足，则合一失败。这样就排除了相当一部分由于搭配不当所造成的歧义。

四、实验结果和结论

我们对于3000个覆盖了现代汉语主要句型的汉语句子封闭集进行了调试，调试后翻译的正确率达到90%左右。目前我们所使用的汉语分析规则库中的规则不超过400条，词典中约有4000个汉语词条。

我们下一阶段将要进行的一项主要工作是词典的扩充，争取使之成为一个实用的系统。另一项工作是尝试引进语料库和统计的做法来帮助获取知识以提高翻译的准确率和效率。

参考文献：

[1] 段慧明，俞士汶，机器翻译评测报告，《计算机世界》报1996年3月25日，第183页

[2] 俞士汶等，《现代汉语语法信息词典》规格说明书，中文信息学报，第10卷，第2期，第1-22页，1996年

[3] 冯志伟，自然语言机器翻译新论，语文出版社，1995

[4] Martin Kay, Unification in Grammar, In V.Dahi & P.Saint-Dizier(Ed.) Natural Language Understanding and Logic Programming, North Holland, Amsterdam, The Netherlands.

[5] 沈阳，郑定欧主编，《现代汉语配价语法研究》，北京大学出版社，1995

[6] Gazdar G., Mellish C., Natural Language Processing in Lisp, Addison-Welsley Publishing Company, 1989

[7] 俞士汶，关于计算语言学的若干研究，语言与文字应用，1993年第3期

[8] 刘群，张祥，基于软件构件的机器翻译研究方法，待发表

[9] 常宝宝，詹卫东，一个汉英机器翻译系统中的语义处理框架，待发表

[10] 詹卫东，刘群，语义分类在汉英机器翻译中的作用及其存在的问题，待发表

致谢：

本课题是在中国科学院计算技术研究所张祥研究员领导下开展的，并得到了北京大学计算语言学研究所俞士汶教授的悉心指导和大力支持。前后参加过本课题并作出重要贡献的还有北京大学计算语言学研究所的周强、陶晓鹏和中国科学院计算技术研究所的叶煜、王斌等同学和同事。在此一并向他们表示衷心的感谢。

* 本项目的研究受到863-306资助，合同号为863-306-03-06-2