基于动词配价模式日汉机器翻译系统的设计和实现

孙勇、陈群秀

（清华大学智能技术与系统国家重点实验室）

摘　要：实用性对翻译系统的翻译速度和译文质量提出了较高的要求。本文采用基于动词配价语法、辅以格语法和语义分类的综合语言模型，应用于实用型的日汉机器翻译系统中，对减少分析过程中的歧义、提高翻译速度和译文质量起到了较好的效果。

关键词：日汉机器翻译　配价模式　惯用型匹配　捆绑

Design and realization of a verb-valence based Japanese-to-Chinese machine translation system

Sun Yong, Chen QunXiu

(Intelligent Technologe and Systems Lab.of Tsinghua university)

Abstract: Practicality requires higher speed of translation and better translation result. This paper uses a integrated language model which is based on verb-valence grammar and case grammar、 semantic categorization, in a practical Japanese-to-Chinese machine translation system, it is favorable for decreasing ambiguity in analysis process, increasing translation speed, and generation better translation result.

Key words: Japanese-to-Chinese MT, verb-valence, idioms match, grouping

一、引言

二十世纪八十年代，机器翻译研究开始进入了全面发展的黄金时代，但是也面临着一些严峻的问题。例如，机器翻译系统的译准率长期徘徊在70％左右，译文的可读性、系统的鲁棒性尤其是开放性都不尽人意[1]。社会迫切需要对真实文本进行大规模的语言信息处理[2]，而过去几十年由于种种条件限制，研制的自然语言理解系统和机器翻译系统同当今社会对大规模真实文本处理的期望甚远。各国的计算语言学的学者和科研工作人员正在为实现真实文本处理而努力，为此采用了各种模型、多种处理策略和手段。

机器翻译系统的译准率和译文可读性之所以不尽人意，是因为在翻译过程中存在着下列问题：1. 源文句子分析时的语法结构和语义结构存在歧义；2. 多义词译词选择问题；3. 译文生成时还存在介词或助词的多义选择问题。另外，实用化的机器翻译系统还面临一个分析的准确性、存贮空间和分析速度综合考虑甚至于折衷考虑取舍问题。

清华大学计算机系近几年来在探讨日汉机器翻译的研究中，根据对几种机器翻译系统研究方法的体会，设计了一个基于配价语法、辅以格语法和语义类型的综合语言模型，在此模型基础上实现了一个日汉机器翻译系统分析器并在向实用化方向努力.

二、基于配价语法的综合模型

在自然语言理解中，使用依存语法、格语法的情况不少。配价语法就是从依存语法发展而来的，属于支配关系语法体系。支配关系语法的最大的共同点是以动词为句子核心，其他成分则受动词支配，与动词挂钩。法国人 L. Tesniere 首先提出了德语的配价语法，而日语由于其语言特点也很适合用配价语法（又称结合价语法）来进行分析[3]。从配价语法看句子构造时就成为处理以动词为中心的句子构造类型。“配价”是把化学上某个原子和其他几个原子结合的思考方法应用到语言学上来。所谓配价与原则参数语法中的“论元”是一个意思，即指一个动词与几个名词或词组（共演成分）的支配－从属关系。例如，汉语的“下雨”是0价动词，“跑”是一价动词，“吃”是二价动词，“给”是三价动词.

配价语法认为，每个动词都有其一个或若干个与配价有关的特定句型。日语由于其语言特点（动词价信息较清楚，有に、で、へ、は等助词）很适合用配价语法来处理。格语法在考虑深层意义关系方面有特色，而配价语法擅长处理表层句法结构，把这两者结合起来可构筑对日汉机器翻译有用的语法，再加上将体词进行语义分类，就能解决动词多义选择，助词多义选择、甚至某些体词同形多义、同音异形问题。另外，还要加上动词的时态、语态、体、语气等语法、语用信息等综合表示。我们的日汉机器翻译系统中采用的就是这样一个综合表示模型。我们采用的语义分类体系是在借鉴国内外若干个分类体系基础上制定的一个语义分类体系[4][5].

我们选择动词配价语法作为语言模型的基础，是因为它具体而详尽地规定了动词的使用规则，而对句子中的其它词则是根据语义进行分类。由此可见，动词配价语法既在很大程度上避免了因为语法过于粗泛而导致在分析中产生大量的歧义，也避免了对每个词都做具体分析而产生的系统资源不足的问题。因此，它是一种折衷、切实可行的方案。

例如，动词“案内する”有以下几个配价模式（关于配价模式中符号的约定请参看有关资料）：

1. {人} {は|が} {场所} を V
2. {人} {は|が} {人} を V
3. {人|集体} を {事|精神活动} に V

以下句子依次对应于上述配价模式:

1. 原文: 息子が駅まで道を案内する。 译文: 儿子陪同游览去车站的道路。
2. 原文: 私は先生を案内します。 译文: 我陪先生游览。
3. 原文: お客様を特别ヤールにご案内する。译文: 为客人介绍特别大减价。

本系统利用格语法和语义分类辅助动词配价模式，更明确地定义了句子的结构。通过格框架模式可以给出确定的分析结果。

本系统采用的语言模型的优点可以归纳为以下四点:

A. 减少了分析过程中的歧义

因为动词配价模式直接清楚地表示了句子结构，所以减少了分析过程中的歧义。若有两个句子的语法结构一致，但是语义结构不同，则根据句中单词不同的语义分类可以找到与这两个句子对应的不同的配价模式、主动词译词和句子译文模式。

B. 减少了生成过程的歧义

由于语言模型的清晰性，所以减少了生成过程的歧义，同时也有利于提高生成译文的速度和准确度。

C. 动词配价模式有利于提高译文的质量

一个动词有多个配价模式，每个配价模式都可以定义自己的动词译文。这样就可以比较准确地翻译动词。在翻译的过程中，可以根据不同的匹配生成比较恰当的动词译文。

因为动词的多个配价模式之间都是相互独立的，每个配价模式都可以定义自己的生成译文的方式，所以与配价模式对应的汉语译文也就可以比较贴切，并有利于独立地调整每个动词的配价模式，而不会对其它动词的翻译产生副作用。

D. 使用动词配价模式有利于提高翻译的速度

利用捆绑可以把较长的句子划分为若干个较短的组成部分。对句子整形后，可以快速地判断句子是否与动词模式匹配。因此有利于提高翻译速度.

本系统采用的语言模型既然有上述长处，也必然有某些缺点，其缺点及我们采取的相应对策如下:

A. 配价模式的数量巨大

因为每个动词都有自己的配价模式，而且通常都有多个。所以动词配价模式的总数量必然伴随着系统的开发以相当快的速度增加，由此导致的主要问题是动词配价模式的查找与匹配可能会花费较多的时间。

我们可以采取一定的措施来抵销这一缺点引起的过多时间开销，提高系统的翻译速度，如建立一级、二级索引来加快动词配价模式的查找，加快句子与配价模式的匹配速度等。

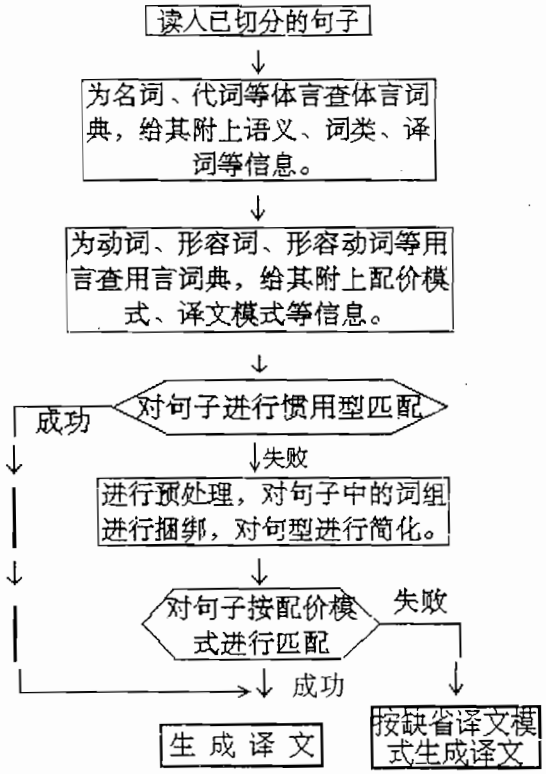
B. 系统的适用范围受到制约

“配价模式+格框架模式”的组合在具有清晰性的同时，也限制了其自身对语言现象的概括能力。要正确地进行翻译，系统辞典内必须事先有恰当的、能够与句子匹配的模式。因此，为了保证系统的实用性，一方面可以增加配价模式的数量，以此确保用户输入的句子可以得到正确的翻译；另一方面可辅以其他分析手段，我们就采用基于实例的翻译方法、作为对基于配价语法的后续系统。

三、系统实现

A. 系统流程

经过对系统的深入分析，决定日汉翻译系统的流程如下:



B. 惯用型匹配

惯用型的匹配包括两个部分，一部分是日常用语（即简单惯用型）的匹配，一部分是分立式惯用型的匹配。

在日语文章和日文对话中有许多重复出现的日常用语（即简单惯用型），因此对简单惯用型进行处理是翻译过程中重要的一步；对句子进行简单惯用型匹配的一个特点是不论句子形态如何，将整个句子直接与惯用型辞典中的简单惯用型进行匹配，如匹配成功，则认为此句子是简单惯用型，直接得到其译文。

简单惯用型举例:

おはようございます。早上好。　しつれいします。失陪了。

いらっしゃいませ。欢迎光临。　失礼します。　　失陪了。

分立式惯用型可当作复杂句进行处理。

C. 利用规则对句子进行捆绑

惯用型匹配之后就是对句子进行预处理捆绑。预处理的规则放在规则库中。在进行捆绑时，依次扫描句子中的单词，按其词类取出相关规则依次匹配句中的以后单词。

若是复杂句，还需按动词配价模式进行捆绑。

捆绑完毕，对句型进行简化，之后就可以进行配价模式匹配了。

在较长远的考虑中，还可以从与句子匹配的配价模式拉出语义关系模式代号，同时进行同音异形词和多义词辩识以及格助词意义确认，记下谓语的时态、体态，句子的语态、语气等信息，为以后结合句子的语义信息生成更佳译文作准备。

捆绑规则举例:

1. 名 名 M {1} {2}　　2. 体言 の 体言 M {1}的{3}

3. 形容 体言 [2] {1}的{2}　　4. 动 3 体言 Y {1} {2}

5. 体言 や 体言 (など) [3] {1}和{3}等

在规则中: Y 表示需按动词进行匹配; M 表示以最后一个词为主词; [n]表示第n部分为主词; 缺省主词为最后一个词; 最后为该规则的译文模式。

D. 捆绑与结构树的设计

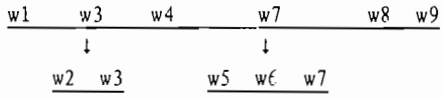
捆绑是翻译过程中对句子进行预处理的重要过程，在此过程中用规则判断句子中的修饰关系、并列关系以及扩展成分等，并对以上关系捆绑成词组，之后可对句子的主体整形和简化，句子与配价模式进行匹配时仅使用句子主体即可，使过程趋于简单明了。

然而，如何很好地表示捆绑结果，并保留捆绑过程中的信息，以及便于生成译文？从语义树的表示中得到启示，系统中设计了一个名为结构树(Struct Tree)的数据结构；其基本思想是结构树的上层以句子的主词表示句子的主体，若上层的某个词是捆绑后的主词，则这个词指向一棵子树，子树中存储了捆绑信息，包括该子树的译文模式；此定义是递归的。

例: 若结构树的最初形态为:

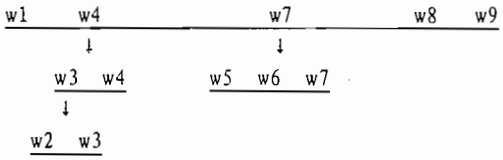
w1 w2 w3 w4 w5 w6 w7 w8 w9　(w表示一个词)

如将第2、3个词和第5、6、7个词捆绑到一块，主词分别为第3和第7个词，则结构树变为:



句子结构树的主干里不再有W2、W5和W6，而W3、W7分别指向一颗子树。

如再将第3、4个词捆绑到一起，主词为第4个词，则结构树调整为:



结构树可有多级子树，并且每级子树都包含了相应译文模式。

E、动词配价模式匹配和译文生成

经过捆绑，对句子整形后，动词配价模式的过程就变得相对简单了。从理论上讲，只要将句子的各个主词与动词配价模式的各条件进行匹配即可，其主要难点是要处理好配价模式中可选项是否保留的问题。

在捆绑过程中，配价模式中的某个条件是否与句子中某词匹配的信息也存贮在结构树中。生成译文的过程是对结构树的递归访问，在此过程中利用了结构树中各级子树的译文模式。从系统测试过程中可看出，生成的译文都符合设置译文模式的设想。

四、生成译文过程中的几个问题

整个日汉翻译系统是一项大工程，本分析子系统只是其中的一部分，它接在分词系统之后，而在其之后还有基于实例的翻译系统和使用断段分析方法的翻译系统，而且以上各系统不是独立的，它们之间可能由主控程序控制相互调用，即在翻译某个句子时，可能使用不仅一种翻译方法。

A. 句子结构歧义

对句子进行捆绑时应利用单词的语义信息，否则不能解决句子的结构歧义的问题，可能出现误捆绑。

进行捆绑时，不仅要判断句子是否符合规则提出的词性要求，还要考虑句子中单词的语义信息。如对规则“名词 + 名词”，当句子中两个相邻的名词具有相似的语义类，即其相似度较高时，这两个名词可以捆绑到一起；然而，如果这两个名词语义相差很大，则不应对其进行捆绑。例如对句子的部分切分“中国 N 北京 N”就可以捆绑为一个整体，而“今年 N 北京 N”则不应捆绑；若进行捆绑，则是误捆绑，从而导致进行动词配价模式匹配时，对应匹配成功的模式匹配失败。对于“名词 + 名词 + 名词”以及“名词1 の 名词2 と 名词3”（这是定语修饰关系与并列关系混合的例子）等规则，同样要解决上述问题。

对以上方法不能辨别的歧义问题，可以采用模糊处理的方法来解决。

B. 循语义树匹配

在进行动词配价模式匹配时，对于模式中条件为语义类时，当单词的语义类若是条件语义类的一个子类时，也应匹配成功，即在进行匹配时应进行循语义树的匹配。

在系统实现中可以利用现有语义树文件，该文件的格式为：语义类名称 语义类编号

语义类编号形如：a1.a2.a3.a4.…，表示该语义类所属的大类及子类编号。例如：非生物 2.1.2；天然物 2.1.2.1；化合物 2.1.2.1.6 等等。

在进行动词配价模式匹配时，分别读出条件语义类的编号和单词语义类的编号，若单词语义类编号的某个前缀与条件语义类相同，则表示为其下位可以由继承性而匹配，否则不匹配。

C. 译文中助词的生成

译文的生成是一个复杂的过程；而且，汉语和日语分属不同的语系，表达方式差别较大。因此，生成汉语译文时需考虑诸多的因素，才能生成通顺、流畅的译文。

在系统测试过程中，较为频繁和突出的问题是“的”的生成问题。在设置和利用译文模式时，如果“的”的生成不当，将会大大影响译文质量。经过考察“的”的使用规律，可采用如下处理方法：

1. 形容词 + 名词：生成时加“的”；2. の + 名词：一般生成时加“的”；3. 代名词 + の + [human]：不生成“的”；4. 形容动词 + な + 名词：生成时加“的”，等等。

对译文中“着”、“了”等词生成也要作相应考虑。

五、结束语

过几年的努力，我们在基于动词配价模式的日汉翻译系统的开发中取得了一定的进展，获得了一些实践经验，并用此翻译系统得到良好的译文。然而，日汉机器翻译涉及学科领域广、研制周期长，我们还有很多工作要作：

1. 翻译理论的提高和系统知识的扩展；
2. 整个日汉翻译系统的联接和各部分的合作；
3. 机器词典的扩充和管理；
4. 较复杂句的翻译；等等。

参考文献

[1]. 冯昶，程光远，陈群秀，黄昌宁，“一个基于直接转换的日汉机器翻译实验系统JCBDT-1”，《中文信息处理第二届国际会议论文集》，1992年10月，P280－286。

[2]. 黄昌宁，“关于处理大规模真实文本的谈话”，《语言文字应用》杂志，1993年第2期，P1－10。

[3]. 水谷静夫，石绵敏雄，荻野孝野，草雉裕，《文法と意味》，朝仓书店，1983年9月1日初版。

[4]. 陈群秀，张普，“信息处理用现代汉语语义分类体系（之一）：属性分类”，《计算语言学研究与应用》，北京语言学院出版社，1993年10月。

[5]. 陈群秀，“有关语义分类体系研究的几个问题”，1992年，92全国机器翻译学术会议论文集《机器翻译研究进展》。

[6]. 冯昶，“一个实用型日汉机器翻译系统的研究和初步实现”，1994年，硕士论文。

[7]. 黄河燕、陈爱萍，《我国机器翻译的技术和产品现状》，《计算机世界》，1994年2月。

[8]. 陈群秀、李咏玖，“日汉机译系统中有关汉语生成的几个问题及处理方法”，1991年全国计算机语言学联合学术会议论文，1991年11月。

[9]. 仁田义雄，《语汇论的统语论》（日文），1980年，明治书院。

[10]. Hutchins, John, 《Latest Development in Machine Translation Technology》, 1993, 《The Fourth Machine Translation Summit》 Proceedings;

[11]. Vasconcelos, Muriel, 《The Present State of Machine Transiation Usage Technology》, 1993, 《The Fourth Machine Translation Summit》 Proceedings;

[12]. 张延群，《模糊假言推理中翻译规则的分析和比较》，《计算机科学》杂志，1993年，Vol20. No. 4;