雅信 CAT 和东方快车机器翻译软件的分析及建议

刘彬　谭外元

（中南大学外国语学院，长沙，410075）

E-mail：olympcr2000@sohu.com

摘要：本文叙述了雅信 CAT-2.5 简体英汉双向版（网络版）和东方快车 3000 英汉翻译软件对源语的切分以及由源语向目标语转换的过程。认为使用上下文无关语法，并把词作为转换单位的程序设计，不能得到较理想的译文。建议以动词为中心的小句为切分单位，采用依存语法，以提高译文质量。

关键词：机器翻译；东方快车 3000；雅信 CAT；切分；转换单位；依存语法

引言

加入世界贸易组织之后，中国对外交流更加频繁，机器翻译软件在帮助人们克服彼此间的语言障碍，增进相互了解方面起到重要作用，发挥着省时省力、快速简便的优势。目前机器翻译处于第三个繁荣时期，机译研究走向了实用化，出现了大批实用性较强的系统程序，机译产品开始进入市场，变成了商品，逐步完成由实用化向商品化的转换过程。

1. 分析：

1.1简介

雅信CAT-2.5以词为单位进行切分，东方快车3000以词组为单位，两种软件都带有专业词库，并综合了近年来计算语言学的一些成果，如引入了复杂特征集等，对于宾语从句，定语从句翻译较好。国内的翻译软件似乎基本上采用上下文无关语法，这使编译程序过程用时短，适于短时开发，这种语法60年代曾在国外机译研究中被广泛采用，但它不足之处在两种软件中也有所体现。以下是从两本语法书中挑选的例句作为源语，通过两种软件翻译，分析它们的运行情况。

1.2词典

1.2.1概述

词典从功能上大致可分为固定词典和综合词典。雅信CAT系统具有某种处理习语的机制，词典在词条中规定了一些固定词组，翻译句子时，软件针对每一单词都有备选词条。例如源语中有talk一词时，软件能自动罗列talk about(谈论)，talk with(与……交谈)，talk around(兜圈子/说服)，talk down(驳倒)等41个中文词条；而东方快车3000则附带东方快典备查以解决一词多义问题。两种翻译软件实际上都采用语料库的方法解决歧义问题。两种软件能很好地翻译一些例句：

例1. Bob and Jack are twin brothers.

东方快车3000译文：鲍勃和杰克是成双的兄弟。

雅信CAT译文：鲍勃和杰克是孪生兄弟。

人工译文：鲍勃和杰克是孪生兄弟。

例2. What destroyed the church?

东方快车3000译文：什么破坏了教堂?

雅信CAT译文：什么毁坏了那教堂?

人工译文：什么把教堂毁坏了？

1.2.2 固定词组

许多固定词组未能译出是由于习语库规模不够大，资料不够齐全，未采用开放式语料库，并存在使用语法分解处理固定词组,使译文背离原文意义的问题。固定词组通常只能作为一个整体单位翻译，而不能分解处理。基于这一点，上述两软件似乎可以考虑建立更加完善的语料库。

例3. He went off, gun in the hand.

东方快车3000译文：他爆炸了，在手里的枪。

雅信CAT译文：他去off，gun 在hand。

人工译文：他离去，手里拿着枪。

例4. He said he had bumped into the young man the day before.

东方快车3000译文：他说他前一天闯入了年轻的男人。

雅信CAT译文：他说他已经击年轻人人前天。

人工译文：他说他在前一天偶然碰到了那个年轻人。

上述两例表明两种软件习语库具有一定规模，但由于翻译程序作分步翻译之前未能搜寻习语库，或者因为习语库本身储存较少，所以有时在确定习语意义时，不能作出正确选择。

1．2．3 综合词典

综合词典是翻译软件的关键部分，包括词汇、语法、语义、转换规则。像其它许多翻译软件一样，上述两种软件都引入了复杂特征集（Complex Feature Set），用于描写各个词条，包含有多个特征项的语法、语用、语义信息，以及译文信息（机译系统所需）等等。引入这一概念，是因为语法树的每个节点实际上包含着丰富的内部结构，或者节点本身一系列属性的描述，或者节点所包括的下层节点描述，使用单个符号无法表示这些内容，其信息量太少，因此也就无法实现对语言的全面分析和理解。目前词典中的句法语义信息可以大致分为三类，在不同程度上体现了功能思想。（一） 基本信息：一个词所属词类、语义类。这是对该词的功能作最一般性的概括。（二） 搭配信息：一个词跟其它成分的组合能力。这包括句法和语义两方面。（三） 位置信息：一个词充当句法成分的能力。

1.3 歧义处理

语言中的同形歧义既反映在单词上，又反映在由单词组成的各种结构上，形成词汇歧义(Lexical Ambiguity)和结构歧义(Structural Ambiguity)。两种软件对于歧义现象的处理都表现出一定能力：

例5. When can you finish writing the letter?(letter词汇歧义)

东方快车3000译文：什么时候你能完成写信?

人工译文：你什么时候能写完这封信?

例6. Hope to hear from you soon. (hear from 结构歧义)

雅信CAT译文:希望不久收到你的来信.

人工译文:希望早日来信.

1.3.1词汇歧义

词汇歧义指的是单词的词义存在两种以上的解释，使软件翻译往往出现不正确的译本，这是由于软件以词为切分单位，使译文与原文在词层上一一对应所造成的。

例7. China is the largest country in Asia.

雅信CAT译文：瓷器是那大的国家在亚洲。

人工译文：中国是亚洲最大的国家。

例8. All things considered, I think I ought to award the job to Smith.

东方快车3000译文：所有的事情考虑了，我认为我应该授于工作到铁匠。

雅信CAT译文：万事considered，I认为我应当将那工作赏给铁匠。

人工译文：既然各方面都考虑到啦，我想我该把任务交给史密斯了。

例9. Shanghai stands on the Huangpu River.

东方快车3000译文：上海承受在……之上huangpu河。

雅信CAT译文：上海坚持那huangpu 河。

人工译文：上海位于黄浦江畔。

上面歧义的词组分别是："China, award the job to, stand on …River "，当前，自然语言处理过程普遍采取两种歧义消解方法：制约(Constraint)，优选(Preference)。所谓制约法，就是利用句法，语义制约条件，排除不能满足制约条件的结构，从而达到歧义消解的目的。所谓优选，就是在若干个存在歧义的候补结构中，选出一个最优的结构，从而消解了歧义。对于例7，8，9出现的歧义现象可用制约方法来解决。

1.3.2结构歧义

结构歧义分为附属歧义（Attachment Ambiguity），空位歧义（Gap and Filling Ambiguity），结构功能歧义（Analytical Ambiguity），我们将在下面进行具体分析：

a. 附属歧义

附属歧义是指某一句法结构没被正确判断为另一句法结构的修饰语而造成的歧义。

例10. Most of the products on display are new ones.

东方快车3000译文：在展出的产品的大多数是新的。

雅信CAT译文：大多数那产品陈列是新的东西。

人工译文：展出的大多是新产品。

例11. He always considers himself in the right.

东方快车3000译文：他总是在权利考虑他自己。

雅信CAT译文：他总是认为他自己有理。

人工译文：他总以为自己是正确的。

例10中，介词词组on display 修饰the products，东方快车较雅信处理得好些，但两种软件对于new的解释有偏离，应该是指技术上的新颖，而非外观上的簇新。而在例11中，介词词组in the right是做补语，雅信CAT的译文是比较正确的。

b. 空位歧义

这是由于某一句法成分移动后留下深层意义上的空位(Trace)而引起的歧义。

例12. I am not the talent you thought me.

东方快车3000译文：我不是你认为的才能我。

雅信CAT译文：我是非那才能你认为我。

人工译文：我不是你所想象的那种天才。

空位歧义是由于两种软件采取的上下文无关语法多从句法线性（Linearity）角度，而未从语义层次（Hierarchy）上分析其深层意义而产生的。如：例13中两种软件都没有正确处理think后接名词作补语的情况，因此未能正确译出"将……想象成……"的意思。

C. 结构功能歧义

这是因为机译软件不能正确判断语法结构功能而引起的歧义现象。

例13. He admitted to the colleagues that he had done it without the president' s permission.

东方快车3000译文：他承认了他没有总统的允许做了它到同事。

雅信CAT译文：他向那同事承认他结束它没有总统的许可。

人工译文：他向同事们承认他做此事未得到校长的同意。

例14. She explained to her friends that she had made the mistake chiefly out of carelessness.

东方快车3000译文：她向朋友解释了她主要从粗心犯了错误。

雅信CAT译文：她向...解释她朋友那她有制造那弄错主要地粗心的无法达到。

人工译文：她向朋友们解释她犯这错误主要是由于粗心。

在例13中，东方快车3000对于that引导的小句功能判断不正确，它应作为主动词admit的补足语，而“to the colleagues”是主动词作用的对象；在例14中，雅信CAT也没有正确认识到that引导的小句是主动词explain的补足语，“to her friends”是主动词作用的对象。这说明对于结构功能歧义的解决不能单从句法结构上入手，而要进一步考虑到语义因素和限制。

1.4 转换

雅信CAT从这种软件的切分步骤来看，很明显以词为单位，在词层上体现从源语到目标语的对应，这一点可从下列句子中清楚地体会到：

例15. Although hardworking, he could not earn enough even to support himself.

雅信CAT译文：虽然hardworking，he couldn’t 赚的充足的均匀的到支持自己。

人工译文：尽管他努力工作，收入甚至不够自己糊口。

因此，以词为单位的转换很难对源语作出完整的正确分析，例15与人工译文相差甚远，软件不能识别enough…to的固定句型；而且由于机译以词为单位进行切分，把副词enough译成了形容词“充足的”，even翻成了“均匀的”，support也脱离语境译为“支持”。这些例子说明该软件采取词层对译，未形成对文本处理的整体概念，同时也忽略对语用，语义，句法等各因素的分析。上下文无关语法在切分和消歧两方面还存在明显缺陷，譬如在双语转换中，切分单位过小，造成语义缺乏连贯，并且后处理程序空缺，这样难以译出合乎汉语表达习惯的句子。

例16. Mary is always the first to come and the last to leave.

东方快车3000译文：玛丽总是来的第一个和离开的后面。

雅信CAT译文：玛丽总是第一个未来地和上次丢弃。

人工译文：玛丽总是第一个来，最后一个离开。

两种软件未能识别不定式作定语的英语结构，也就未能译出合乎中文习惯的句子，由此可见软件在运行语料库处理固定词组及其采用的上下文无关语法还存在问题。笔者认为源语与目标语的对照应存在于两个语义丛之间，而并非词层的一一对应。词汇意义应该是由词与词之间的关系所决定，无论源语还是译语，句子中的词义都应该是上下文依存的。

雅信CAT的处理方式是先列出单个词语的意义，将其转换成中间语言（Interlingua），分析并运用源语的语法结构对单词对应的意义进行排列，输出译文句子；译本未引入译语的语法结构来表达汉语语句特点.

例17. Do you like to play tennis or football?

雅信CAT译文：你喜欢打网球或足球吗?

人工译文：你喜欢打网球还是踢足球?

而东方快车3000则是以句为翻译单位，与雅信译文相比，有其优点。

例18. The French people are very friendly to foreign tourists.

东方快车3000译文：法国的人对外国的旅游者很友好。

雅信CAT译文：法国人人是莫逆到外国游客。

人工译文： 法国人民对外国旅游者非常友好。

2. 建议

2.1 语法

2.1.1上下文无关语法

上下文无关语法(Harrison, 1978;Hopcraft and Ullman, 1979)是一个元组G=(V_N, V_T, P, S), 其中，V_N是非终结符集，V_T是终结符集，V_N, V_T均为有限集合，并且$V_N \cap V_T = \Phi$。$S \in V_N$是起始符。P是一个产生式有限集合，P中的产生式形式为：$A \rightarrow \alpha$。此处，$A \in V_N$，（A为单个非终结符），$\alpha \in (V_N \cup V_T)$。即当用$\alpha$替换A时，与A的上下文环境无关。如果集合P中的每个产生式都有形式$A \rightarrow BC$或$A \rightarrow \alpha$，因为A，B，C $\in V_N$，$\alpha \in V_T$，实质上这就是乔姆斯基文法分类中的2型文法。

2.1.2依存语法

法国语言学家吕西安·特思尼耶尔（Lucien Tesnière）在1959年《句法结构基础》一书中具体提出配价（依存）语法概念，并指出“联系（La Connexion）、组合(La Jonction)和转移(La Translation)是概括一切句法结构现象的核心”。它是一种结构语法，主要以研究谓词为中心而构句时由深层语义结构映现为表层句法结构的状况及条件，谓词与体词之间的同现关系，并据此划分谓词的次类。

依存语法重视描写句子中词与词之间的各种关系。特思尼耶尔也明确提出，句法分析的任务在于研究句子，而句子不只是词的简单组合，它是包含着词与词之间多种关系的“有组织的整体”，他把关系视为和词同等重要的句子构成成分。依存语法用于句法分析中具有以下五个主要特点：1.注重成分之间的外部联系，强调了各成分之间存在的功能关系；2.中心词驱动，突出了中心词在句法语义上极其重要的作用，为深化分析创造条件；3.表达简洁，便于计算机处理，易于计算机获取知识；4.因为得到的从属树层次不多，结点数目少，可清晰地表示句子中各单词之间的依存关系，在计算机对句子进行句法分析时，简化分析短语结构过程中所需的众多环节，使分析工作简单明了，而且搜索空间较小；5.不以正确划分词类为语法描述对象，而直接把动词作为对象，绕开了给汉语定词类的棘手问题，这同使用传统的词类和短语规则对语言现象进行描写时知识颗粒过大的情况形成了鲜明的对比。依存关系体系既可应用于基于规则推理和语料库统计知识的句法分析器，也可以用于对语料库的句法标注。具体以图1, 图2为例, 两图均采用中序排列，并用二叉树（Binary Tree）方式进行结构分析，图2中动词*is*, *water*, *am*分别作为三个节点，并根据其左右子树，将原句分为： The problem is, who will water

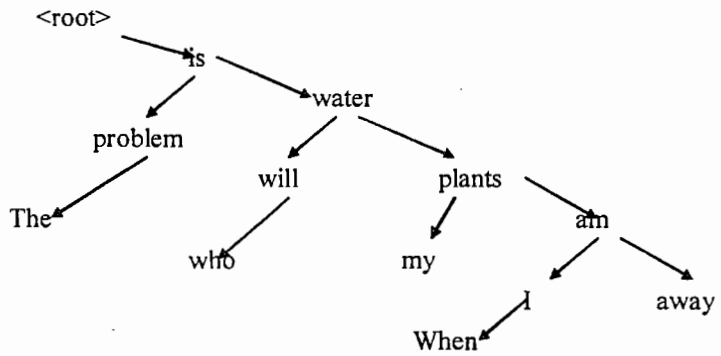

图 1.依存语法英语源语分析

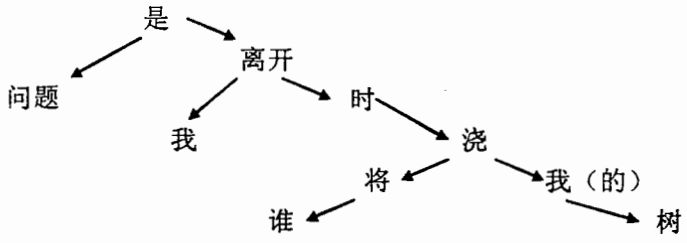

图 2. 依存语法汉语译文分析

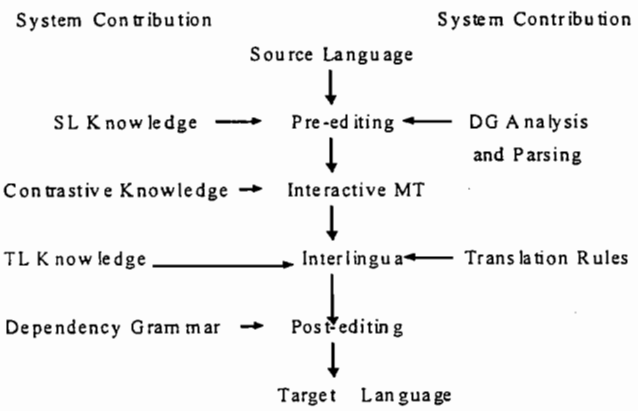

图 3. 翻译流程图

my plants, when I am away三个部分；而图3中对应以“是，浇，离开”三个动词为中心，将译文句子同样分成三部分：问题是，我离开时，谁将浇我的树。原文和译文对照很工整，首先原文形成一个判断句型，两个词组：water plants, be away,运行时，软件先转换这三个核心部分，然后再将修饰成分围绕已生成的三个核心部分，按照目标语语序规则排列，如附属语（定语、状语）前置。词义的选择应在具体语境中进行，其意义受到小句中其它词项的限制。这样做能有效地处理机译中的歧义问题，简化机译运行的中间环节，提高译文的质量。

2.2机译的分工步骤

目前，机译一般分为三个部分：（前处理）Pre-editing，（交互式机译）Interactive MT，（后处理）Post-editing。在pre-editing 中,计算机对源语进行分析和切分，这一步的关键就在于一种受限制语言的语法和语义分析能力，而这种语言能保证机器内置的知识准确无误地运行。尤其是在交互式机译过程中，许多翻译系统采取源语和目标语两个窗口相对照的方式，配以在线双语词汇库，交互式地逐句进行翻译，通常在从语料库选择译文的过程中暂时停顿。交互式翻译可减少中涉及的语篇碎片。后处理是直接影响译文质量的关键步骤，东方快车3000将后两步合二为一，而雅信CAT配有大量的备选词汇，让译者使用语料库来完成后处理工作。源语和译文之间还存在着中间语言。我们必须清楚，目标语应是一种合乎语法标准规范的语句是不能和机器语言等同起来的。我们认为比较合理的翻译形式可用图表示为：（见图3）

3. 结论：

通过上面对于东方快车和雅信软件的对比分析，我们认为两种软件已在源语分析和译文生成上取得了很大成绩，如果更进一步地采用以动词为中心的小句为切分、转换单位并引入依存语法为语言生成模式，内嵌翻译规则，考虑到一些语用语义因素，将会使译文效果更加完善。

4. 参考文献：

[1]. 冯志伟. 自然语言机器翻译新论. 北京：语文出版社. 1995年
[2]. 赵铁军. 机器翻译原理. 哈尔滨工业大学出版社. 哈尔滨. 2000年
[3]. 罗选民 谭外元 唐旭日. Matrix 英汉翻译系统的分析及建议. 中国科技翻译. 1999年11月第12卷4期
[4]. 刘海涛. 依存语法和机器翻译. 语言文字应用. 1997年第3期
[5]. 冯志伟. 自然语言处理中的歧义消解方法. 语言文字应用. 1996年第1期
[6]. 冯志伟. 依存语法在机器翻译中的应用. 2001年3月在浙江大学外国语学院的讲稿
[7]. 沈阳 郑定欧. 现代汉语配价语法研究. 北京大学出版社1995
[8]. a. 章振邦. 通用英语语法. 上海外语教育出版社. 1999年7月

b. 张道真. 实用英语语法（第三次修订本）. 商务印书馆. 1992年5月

致谢：原中国翻译协会副会长, 中国中英文化比较研究会创始人, 英汉翻译“信、达、切”标准提出者，著名翻译家刘重德教授对本文进行了悉心指导，在此表示衷心感谢。

作者简介：刘彬（1974-），男，湖南长沙人，硕士生，主要研究方向：依存语法在机译后处理中的应用。
谭外元（1955-），男，湖南茶陵人，硕士生导师，主要研究方向：音系学，语义学。

Analysis and Suggestion on Yaxin CAT and Oriental Express

LIU BIN[1], TAN WAIYUAN[1]

[1](Central South University, Changsha, 410075, China)

E-mail: olymper2000@sohu.com

Abstract: This paper makes an examination on the Yaxin CAT and Oriental Express 3000 machine translation softwares on the parsing of source language and the process from analyzing the source language to transferring to the target language .This paper argues that the software design using the Context-Free Grammar(CFG) and taking the word as the transfer unit limits the analysis on source language and its deep meaning, and suggests that the software take the verb-oriented transfer unit and adopt the Dependency Grammar(DG) to improve the translation quality.

Key words: machine translation;Oriental Express 3000;YaxinCAT-2.5; parsing transfer unit; dependency grammar