

英汉机译系统中的难点分析

郑杰 茅于杭 未末

清华大学自动化系语言研究中心, 北京 100084

摘要: 本文根据开发英汉机器翻译系统的经验, 针对当前英汉翻译中存在的不足和自己在研究中遇到的难点和重要部分进行了分析, 这些难点分布于机器翻译的各个阶段, 对其中的一些难点, 本文探讨了可能的解决办法。

关键词: 机器翻译 自然语言处理

Discussion on Difficulties of English-Chinese Machine Translation

Zheng Jie Mao Yuhang Wei Mo

Speech and Language Processing Research Center, Tsinghua Univ., Beijing 100084, P.R. China

Abstract: In this paper, the shortcomings of the current English-Chinese machine translation system and difficulties of developing a practical machine translation system are analyzed on the author's research experience in this field. These difficulties are scattered in all stages of machine analysis and generation. we discuss the possible solutions to some difficulties.

Keywords Machine Translation, Natural Language Processing

一、引言

近些年来, 机器翻译的研究有了较大的进展, 多种语法体系和方法的产生和运用, 使机器翻译的研究重新焕发出青春。特别值得指出的是, 随着计算机的普及和硬件性能的迅速提高, 已有很多的英汉机器翻译软件进入市场。这的确是一件令人鼓舞的事情, 然而, 现有的实际应用的翻译软件的质量仍难令人满意, 要达到真正的实用化水平, 还有一段艰辛的路要走。

本文对当前各种的应用机译系统的实现情况, 对英汉机译中的难点和重点以及当前应用机译系统中的不足做了较深的探讨。这些难点是存在于英汉机译系统的各个阶段, 任何一个方面处理的不够完善都会使系统的翻译质量下降。我们深深体会到, 在机器翻译中, 句子的分析, 规则的制定以及建立一个适用于机器翻译的词典对于整个机器翻译系统的质量都是非常重要的。

本文结合实际例子, 对这些重点和难点进行了分析, 并对一些问题提出了可能的解决办法。

二、英文语法分析的难点

机器翻译的问题，实际就是找出一个函数，将源语映射到目标语。由于源语和目标语的数量都是无穷多的，因此无法采用枚举的方法。现在的机译系统主要依据的是要素合成原理，这要把原文的句子分解成基本构成要素（原文分析）以及将基本构成要素合成译文句子（译文生成）。只要在反映英语句子结构的语法公式与反映汉语句子结构的语法公式之间建立对应关系，就可以实现两种语言的转换。但在实际的操作中，问题可能会很复杂，例如在英语中作为基本构成要素的名词短语，它的结构有如下的形式：

(英语) 名词短语 → 名词
人称代词
冠词+名词
副词+形容词+名词
名词+介词+名词
名词+关系代词+句子

.....

所以，要正确地得出原文的构成要素，必须对原文进行深入的语法分析。

在机器翻译的系统中，对原文的语法分析可分为词法分析和句法分析两个阶段。下面我们对这两个部分的难点分别进行分析。

2.1 词法分析的难点

词法分析包括单词的形态分析和查找；兼类的判定，特别是名词和动词的兼类；进一步还需要进行译文的选择、兼类处理和译文选择的正确性都直接关系着最后的翻译生成结果。不过这方面已有较多的论述，作者在这里就一些其它方面的难点问题进行讨论，这些问题虽然不大，但也直接影响着译文的可理解性。

2.1.1 量词的添加

当前英汉翻译系统的分析策略都是词→短语→句子[3]。英文的单词具有词形的变化。为我们进行词法分析提供了一定的信息，相对于汉语的词法分析具有一定的优势。但英文缺少量词，因此在词法分析时，数词的处理就需要我们特别注意，我们必须根据不同的情形，加入不同的量词。由于情形的多样和复杂，数词的处理也就不是非常容易。例如以下几个例子：

He is ten.	他是十(岁)。
He did not come until ten.	他十(点)才来。
One of the three is a girl.	三(人)中的一个(个)是女孩。
It is July 2, 1999.	今天是 1999(年)7月2日。

由上面的例子可以看出，如果对数词不做任何处理直译的话，译文是难懂的，而要正确的加上量词，需要对数词常用的用法和待译的句子做出详细的分析，得出数词在句中所代表的意义，才能正确地添加量词。这在一些句子中是很不容易的，例如上面的第二个例句，“ten”既可能时间“十点”，也可能是其它的意思，比如“十岁”，这样让机器区分就更为困难。

2.1.2 定语从句的处理

与汉语不同，在英文中，名词的定语往往是一个后置的从句，而汉语的定语则往往是前置的，例如在下句中：

his house which is very beautiful

如果被译成“非常漂亮的他的房子”，则用户肯定会觉得不象中文。但由于在机器翻译中词总是先合并成词组，然后再对词组和句子进行分析。因此，在上句中，在定语从句未处理之前，“his house”就已合并成一个名词词组，所以定语从句的翻译生成后，就很有可能被处理到物主代词的前边，得到上面这样不太好的结果。这一类的句子在真实文本中是很普遍的，因此处理的粗糙就会直接影响了译文的质量。我们在处理这个问题上，采用了多次分析的策略，即词法分析与句法分析的多次重复进行，当“which is very beautiful”被分析出是“house”的定语从句时，已合并的词组“his house”将会被拆开再重新进行一次包括定语从句在内的词组合并。这样，译文就成为了“他的非常漂亮的房子”，符合中国人的阅读习惯。

2.1.3 具有并列性质的单词和符号的处理

在自然语言处理中，并列词的处理是非常重要也是相当复杂的一环。在英语中，并列连词除了象“and”和“or”等明显的连词外，还有表示并列意义的标点符号，逗号。它们在真实文本中的出现频率是相当高的。因此它们的处理也就直接影响着整体的质量。

并列词可以连接并列的各种词性的词，也可以连接词组和句子。一个好的英汉机译系统必须能够对这些情形做出足够正确的判断和分析。否则，就会使句子无法理解。我们比较一下下面的两个句子。

He is killed by Tom and Jerry.

He is killed by Tom and Jerry is killed by Mike.

在这两句中，“and”的作用是完全不一样的，它在第一句连接两个名词，在第二句中连接两个单句。但如果对“and”的分析不够，或者将对它的处理太过简单化。以上的句子就可能会出现这样的翻译错误：或者第一句被译成“他被汤姆杀死和杰瑞。”；或者第二句被译成“他被汤姆和杰瑞杀死被迈克杀死。”。分析这些错误，对连接词“and”的分析是其中的关键，只要它的并列性质被正确分析出，即在第一句连接两个名词，在第二句中连接两个单句，上面两句是很容易翻译的。在实际情况中，并列词可能连接副词，形容词，名词，动词，名词词组，动词词组，介词词组，句子等，正确区分这些情形需要在词法分析和句法分析中对句子结构做详细分析。特别在连接两个并列词组时，这时连接词的并列性质并不能简单地由其左右的词性决定，因此并列性质的确定尤为困难。

逗号的处理：

逗号也是一种特殊的并列词[1]，它的情形比一般的并列词更为复杂。它除了作为一种并列词存在外，还有各种各样其它的用法，以下的例子可以看出这一点：

He will arrive on July 5, 1998.

He is killed by Tom, Jerry and Mike.

That student is a kind, smart and handsome boy in the class.

Tom, head of the office, orders that everyone must come before ten.

由以上几句中，第一句的逗号是作为时间的一种写法，第二句连接并列名词，第三句连接并列形容词，最后一句是作为同位语的分隔。这几种用法在词法分析中是不可以混淆的。由于逗号的用法的多种多样，要区分这些情形，确定逗号在句中的作用也就不是一件容易的事情。多数的翻译系统仅是简单地把逗号当作子句的分隔标志，把一个句子分成两个子句来分析。

2.1.4 词法分析的其它一些难点

下面我们不再细述，只提出一些容易出错的地方。

- “of”的处理，“of”这个词，既简单，也不简单，如在“three cups of coffee”和“three windows of a house”中，一个是顺序翻译，“三杯咖啡”；一个是倒着翻译，“房子的三个窗户”。在实际中究竟应该如何翻译，须根据前面的词是否作为量词使用。在一些情况下，两者并不容易区分，如这两个词组：“three trucks of stone”（三车石头）和“three trucks of a company”（公司的三辆车）。要区分这两个词组，我们不得不借助于一些先验知识，也就是语义信息来判断。
- “no”否定句的处理，“no”可以是形容词，但在汉语中，并没有相应的否定形容词。因此，需要将其转化为否定句的形式。例如：“He got no pay.”应译为“他没有得到报酬。”而不应译为“他得到没有报酬。”

2.2 句法分析

句法分析可分为复句的分析和单句的分析，同样，句法分析的难点也可以分为单句分析的难点和复句分析的难点两部分。

2.2.1 单句句型的判断

一个好的英汉翻译系统需要对英语的句型做出详细的分类，针对每一类都制定出相应的处理方法。虽然自然语言分析的理论和方法越来越多，我们认为，无论使用什么样的分析和翻译理论和方法，都必须考虑到每一种句型，也就是每一种不同的句子结构。这里作者举出几个不易处理或容易混淆的句型。

“there be”句型：

Once there lived an old fisherman in a village by the sea.

强调句型：

It is not you who stole my wallet.

形式主语句型：

It is not easy to win the game.

形式主语的变形：

The game is not easy to win.

还有许多的须专门加以注意的句型，作者就不再一一列举。在以上几个句型中，我们在句型允许范围内又挑选了容易被忽略的部分句子作为例句。将通常情况下的“there be”改为“there lived”，将强调句与否定句结合起来，但这些都是在实际文本中完全可能会遇到的，实用的系统必须解决好这些问题。

2.2.2 单句语法成分的确定

句子中单词或词组的语法功能的确定是直接关系到译文正确与否的重要问题。对于大多数的句子，我们可以通过对句子中词或词组的词性结构的判断，采用上下文无关语法、转换语法或其它方法做出相应的语法分析。但也有相当一部分的句子仅靠这些表层的结构分析还是不够的。例如下面的两个句子[3]：

I bought a table with three legs.

I bought a table for three dollars.

这两个句子的词性结构组成是完全一样的，但介词词组在第一句中是作为定语，在第二句中则是作为状语。如果介词词组的语法功能分析错误，可能出现这样的译文：“我用三条腿买了一张桌子。”或者“我买了一张三元的桌子。”，第二句虽然还可以使人明白，但这样的错误也是不可以接受的，因为很容易找到由于这种错误导致译文无法读懂的其它情形。从这两个例句可以看出，单纯依靠句子的结构有时是不够的，要对上面两句做出正确的语法分析，我们必须加入系统对语义在一定程度上的理解。

一旦涉及到单词的语义分析，问题就变得非常复杂。从机器翻译的实用角度出发，在我们的英汉机器翻译词典中对名词的语义做出了标注，将所有名词划分成三十个大类，每个大类又根据实际需要划分成一个或多个小类；同时，对所有的动词所接的词义相关的施动与受动名词的词义做了标注在动词中也做了标注。通过这样并不复杂语义标注，基本上已满足了我们的系统中句法分析的需要。

在语义信息应用中，我们根据动词所要求的语义信息或动词本身的特性，还有名词的语义范畴，我们可以对大多数的语法歧义做出判断。以上面的句子为例，为判断介词词组的语法成分，我们需对“buy”，“table”，“leg”，“dollar”这几个词之间的关系进行分析。在词典中，“table”标记为一种具体存在物，“leg”为器官，“dollar”为货币，“buy”为用钱买东西，我们可以建立一些规则来指出这些类型之间的一些内在联系，例如器官是具体存在物的一部分，货币是买东西的交换品，通过这两条规则，我们立即可以排除这两个例句中的语法歧义，得出在第一句中，“with three legs”是“table”的定语，而第二句中，“for three dollars”是“buy”的状语。

由于语义信息的种类繁多，在实际文本中，我们遇到的情形有可能在语义规则库中未指出它们之间的关系。因此，我们需要对真实文本中类似的具有结构上语法歧义的句子进行统计，从而得出各个语义标注项之间的概率统计值。这样，当规则未给出语义项之间的关系时，使用统计值来做出判断。

2.2.3 复句分析中的难点

复句是由单句构成的，通常情况下我们是将复句分解成单句进行分析和处理。但在一些情况下，复句与组成它的单句又必须作为一个整体来考虑，这也是自然语言处理的难点之一。自然语言特点表现为既有一定的规律性，但例外的情形又普遍存在。且在结构上的区别也不明显。乔姆斯基的上下文无关文法在分析人工语言时非常有效，但对于自然语言的处理就显得力不从心了。其原因就在于自然语言是一种描述性的、远非完备的系统，用上下文无关文法也就难以完备地描述。对这些例外情况做出相应的处理，使系统的处理能力能涵盖尽可能多的情形，就应该是实际系统的现实目标。

在一般的情况下，复句中的单句可以从复句中提取出来作为一个完整的部分进行分析和处理，然后在与复句的其它部分进行组合。但在有些情况下，由于英文和中文组句结构的不同，这样做的结果很难使英文译成合适的中文。例如在以下的句子中：

He leaves as soon as I come.

“as soon as”译成中文的关联词为“一...就...”，整句就是“我一来，他就走了。”。由于连词的翻译与子句的翻译生成纠缠在一起，因此能正确处理这种形式的实际系统并不多。现在让我们仔细分析一下这种情形下的译文生成规律，“一”和“就”被分别插入到主句和从句的主语的后边，然后将从句提到主句的前边。这就要求系统在进行子句的分析处理时，必须对子句的语法结构做出明确的分析和标注。我们的系统在总体设计时考虑到这一情形，对子句译文生成采用的是根据句法分析所生成的语法树，重新按照中文句子的语法顺序生成译文。这样，我们就有可能在子句的句法分析完成后，将相应的关联词分别添加到主句和从句的主语后，然后再完成子句译文的生成，这样就正确地将这句话译成了中文。如果完全将子句的分析生成与整句的处理分离开处理，虽然可以将算法简化，但却无法产生正确的译文。同时，我们这种方法也可以处理一般的英文连接词情形。

2.2.4 子句提取中的问题

复句的处理方法是将复句分解成子句进行处理。这样，正确提取子句就成为其中关键的一环。通常情况下，我们可以通过对连接词的分析、谓语动词的分析分解复句，提出子句。但在一些情况下，情形也可以变得很难判断，当子句处于整句的中间位置时，子句的提取的难度也就相应增大。例如下句：

I left my notebook that I need now in the dormitory.

我们需要将子句“that I need now”从整句中分离出来，本句的难点在于最后一个介词词组“to me”的归属问题。人们一看就知道“leave...in...”构成了一个具有插入成分的词组，子句的提取如果考虑到这一点，系统也就能够正确提取出子句，但同时，这也将子句的提取与词组的分析联系起来，无疑大大增加了分析的难度。

2.2.5 句法分析的其它难点

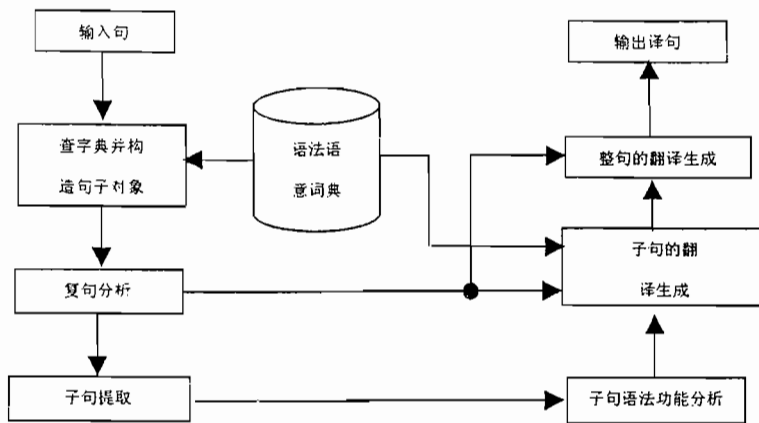
对复句分析中的其它问题，在这里也不再多说，只简单列举一下：

- 连接词省略时，子句的正确提取问题。
- 子句所担任的语法成分的分析问题。子句在复句中的功能有以下几种：作为状语，作为宾语，作为主语，作为定语，还有单纯的并列句。在复句的各个子句被提取出来后，我们还必须对其语法功能做出正确判断，才能正确地翻译出这个句子。
- 子句的生成结果参与整句生成的问题。子句担任的语法功能不同，参与整句的生成的形式也是不相同的。

三、系统的流程

在下图中，我们给出了我们系统的大致处理流程。箭头标出了分析过程中的分析模块之

间的相互影响。



四、小结

本文着重分析了我们在设计和实现一个实用英汉翻译系统时，针对现有的系统的不足和自己在研究中遇到的一些难点，将一些问题提出来，并对其中的一些提出了我们的解决办法。这些问题分布于整个英汉翻译系统的各个方面，它们虽然比较琐碎，但确是我们在实际操作中遇到的，并且我们认为是提高整个英汉机器翻译系统的译文质量中需要注意的地方，也是我们经过一段时间思索的结果。

英汉翻译系统经过这么多年研究，到现在仍然难以真正实用化，生成的译文译准率和质量不高是一个主要原因。我们认为解决的根本办法还要依靠更加深入细致的研究工作。自然语言虽然形式纷繁多样，结构嵌套复杂，描述起来非常困难，但我们认为，语言归根到底还是存在规律的，即使它的规则很多。正因为如此，我们希望能对尽可能多的语言现象进行分析和总结。分析得越细致，解决的问题也就越多。

参考文献

- [1] 张道真：《实用英语语法》，商务印书馆，1986年；
- [2] 俞士汶，朱学锋，“机器翻译导引”，《计算语言学》，1993年7月；
- [3] 刘群，俞士汶，“汉英机器翻译的难点分析”，《1998 中文信息处理国际会议论文集》，507-514，1998年9月；
- [4] 周强，“基于语料库和面向统计学的自然语言处理技术介绍”，《计算语言学文集》，135-147，1996年8月；
- [5] 王嘉欣，基于规则的清华英汉翻译系统的设计与实现[学位论文]，清华大学自动化系，1996年；
- [6] 刘月荣，利用规则消除歧义，提高英汉机器翻译系统的正确率[学位论文]，清华大学外语系，1995年。