汉英机器翻译的难点分析[1]

刘群

中国科学院计算技术研究所 北京 100080

liuqun@mtgroup.ict.ac.cn

俞士汶

北京大学计算语言研究所 北京 100871

yusw@pku.edu.cn

摘 要 汉英机器翻译研究滞后于英汉机器翻译的原因在于汉英机器翻译具有一些特殊的困难。本文根据作者开发汉英机器翻译系统的实际经验，对汉英机器翻译所特有的一些难点，从汉语的语法分析和汉语到英语的转换两个方面进行了较为深入的分析，并对其中的一些难点探讨了可能的解决办法。

关键词 自然语言处理 中文信息处理 机器翻译 算法

Discussion on the Difficulties of Chinese-English Machine Translation

Liu Qun

Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100080

liuqun@mtgroup.ict.ac.cn

Yu Shiwen

Institute of Computational Linguistics,Peking University,Beijing 100080

yusw@pku.edu.cn

Abstract There are special difficulties in Chinese-English Machine Translation. This paper discusses the difficulties in Chinese parsing and Chinese-English transform and gives some proposals.

KeyWords Natural Language Processing, Chinese Information Processing, Machine Translation

[1] 本项目受国家高技术八六三计划资助，课题号为 863-306-03-06-2。

1 引言

近十几年来，机器翻译研究在各方面都取得了很大的发展。多种新型的语法体系的产生，统计方法的运用，使机器翻译研究出现了一个新的高潮；在软件市场上，由于计算机硬件性能的迅速提高和价格的不断下降，使机器翻译软件达到了初步实用的水平。在我国，外汉机器翻译系统，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。然而，汉外机器翻译，特别是汉英机器翻译的研究却进展缓慢，离实用化还有相当的距离。[1]中给出的数据可以说明这一点。从目前的软件市场上也可以看出，比较成熟、达到初步实用化水平的英汉机器翻译系统已有多个，竞争相当激烈，而类似的汉英机器翻译系统却很少，而且几乎还没有一个能达到初步实用的水平。

研究人员普遍认为，汉外机器翻译的难度要远远大于外汉机器翻译。我们从九三年起开始研制一个汉英机器翻译系统[2]，对这一点有着比较深切的体会。在汉英机器翻译中，除了一般机器翻译（如英汉机器翻译）普遍存在的一些困难（如一词多义、结构歧义、语义歧义等等）之外，还存在一些特殊的困难。这些难点分布于翻译的各个阶段，包括汉语的分析和汉语到外语的转换和生成，不过最主要的还是汉语的分析。

关于汉语分析困难的原因，很多人都从不同的角度做过研究，如[6]中就有深入的分析。该文中作者提出计算机分析汉语的特殊困难主要在于以下几个方面：1．汉语同一词类担任多种语法成分且无形态变化；2．汉语句子的构造原则与短语的构造原则基本一致；3．汉语中的虚词；4．汉语的语序；5．汉语的书写习惯。其中前两个特点关系到对汉语语法的全局的认识，而造成这两个特点的根源都在于汉语词类无形态变化。

我们认为，造成汉英机器翻译困难的原因是多方面的。本文以我们的实践经验为基础，从汉语的语法分析和汉语到英语的转换两个方面，对汉英机器翻译所特有的难点进行较为深入的分析，并试图提出可能的解决办法。

2 汉语语法分析的难点

语法分析一般都分为词法分析和句法分析两个阶段。以下我们分别进行分析。然后再探讨造成这些困难的原因以及可能的解决办法。

2.1 词法分析的难点

2.1.1 切分

我国计算语言学界对汉语切词问题研究比较透彻，很多切词系统的正确率都可以达到97%甚至更高。然而，对于机器翻译系统来说，这个问题并不是已经完全解决了。这是因为，机器翻译系统一般是以句子为单位进行处理的，一个句子中只要有一处出现切词错误，整个句子就不可能得到正确的译文。

假设一个切词系统的错误率为2%，在一篇1000词的文章中，大约出现20处切词错误。又假设文章中的句子平均长度为5个词，整篇文章有大约200个句子。那么在这些切

词错误均匀分布（不过分集中）的情况下，这 20 处切词错误就可能导致大约 20 个句子的翻译错误，错误率约为 10%。换句话说，切词阶段的错误率在翻译的过程中将会被“放大”，放大的倍数约等于句子的平均长度。这对翻译正确率的影响是非常大的。

2.1.2 未登录词识别

未登录词不仅汉英机器翻译中存在，其他类型的机器翻译中同样存在。然而对于汉语这种词语之间没有空格分隔的语言来说，还存在一个未登录词的识别问题。困难的主要原因在于，组成汉语未登录词的汉字可能本身又是汉语词。

人类在识别未登录词时主要有两方面：一方面，某几个汉字是否与某一类型的词（如人名、地名等）比较相似，是否符合该类词的一般组成规律，另一方面，如果把这几个汉字当作一个未登录词，是否整个句子会更通顺，更易于理解。现有的这一方面的研究工作多从前一方面来预测可能的某一特定类型的未登录词（如人名、地名、外语音译词等），取得了一些比较好的成果。其实人在理解句子的时候，后一方面的因素同样起着相当重要的作用。但这种判断不仅仅用到了词语方面的知识，更多地用到了句法、语义甚至语境方面的知识，而在计算机自动分析中，未登录词的识别往往处于词法分析阶段，还几乎没有或只引入了极少量的句法和语义知识，因此在这一阶段实现这种判断是非常困难的。

2.1.3 离合词

离合词到底是词还是短语，是个有争论的问题。一种处理方法是，离合词在“合”的时候当作词来处理，而在“离”的时候当作短语来处理。这种方法虽然可行，但总是很勉强的。关键的问题是，离合词即使在分开时仍然是一个整体，而在计算机处理时却只能把离合词的每一部分都当作一个词来处理，如把“打仗”的“仗”，“洗澡”的“澡”字当作名词处理。这样做，不仅不合理，而且会导致分析中很多不必要的歧义组合。

[7]中提出的“语义重心偏移”理论对于离合词问题提出了一种比较合理的解释，不过这一理论在具体的汉语分析中如何操作还有待研究。

2.1.4 语素字

汉语中有很多语素字，它们不是独立的词语，不能单独使用，然而它们的组合能力却很强，很容易用来构成新词或新短语。如“民”字就是一个语素字：“民”字不能单独使用，但却可以出现在 “民心”、“民办企业” 、“国有民营”、“为民请命”、“与民同乐”、“以民为本”、“还政于民”等词或短语中。把语素字作为词来处理显然是不合适的，而如果不作为词，那么对它们构成的新词或短语就无能为力了。另外，汉语中绝大多数单字词同时又是语素字，它们具有很强的构词能力，很容易互相结合组成新词。例如，“冰箱”在港台地区被称为“雪柜”，虽然我们没见过“雪柜”这个词，但我们还是很容易理解它，这是因为，“雪”和“柜”这两个语素的意义是明确的。

2.2 句法分析的难点

2.2.1 短语分析

尽管词的歧义问题很复杂，但毕竟我们对汉语词语的语法特点的认识还是比较清楚

的。在汉语词语的分类问题上，汉语语言学界经过多年的讨论，认识已趋于一致。在北京大学计算语言学研究所开发的《现代汉语语法信息词典》中，更提供了五万多汉语常用词语的详细语法信息。在我们开发的汉英机器翻译系统中使用了该词典，在使用中我们感觉到，该词典中提供的汉语词语的语法信息是比较准确和全面的。

相对而言，我们对汉语短语的语法特点的认识则要少得多。虽然朱德熙先生早就提倡“短语本位”的语法，但实际上我们对短语的认识依然很不全面和清晰。人们很容易以为，只要把词研究清楚了，短语的特点自然也就清楚了。其实不然，词在组成短语后，很多语法特点都发生了变化，短语之间的组合也有一些不同于词语之间组合的规律。而我们对这些规律却知之甚少。

研究短语的语法特点，首要问题同样是分类问题。汉语短语常见的分类方法有两种：一种是按结构划分，分成主谓短语、述宾短语、定中短语、状中短语、联合短语等等，另一种是按功能划分，分成名词性短语、动词性短语、形容词性短语、主谓短语等等。从汉语分析和机器翻译的角度看，按功能分类是更为合理的。这与汉语词语按功能分类的思想也是一脉相承的。然而，具体到汉语短语应该分成哪些类，根据哪些语法功能进行划分最为合理，现在还没有比较好的研究结果。况且仅仅依靠分类并不能说明汉语短语的所有语法特点，同一类短语的语法特点可能还需要足够多的语法属性才能描述清楚。这些问题可能目前还没有引起语言学界足够的重视，但对于机器翻译来说，对于这些问题的研究是非常迫切的。例如，一般认为，“的”字结构都是 np。但我们在调试翻译规则库时发现，如果把所有“的”字结构都当作 np，会出现大量不合理的分析结果，而如果把一般“的”字结构当作 ap，则会好得多。我们的感觉是，真正“的”字结构用作主宾语的情况在实际的文本中并不多见，起码比起“的”字结构作定语的情况要少的多。这至少可以说明，作主宾语并不是“的”字结构的主要用途。再比如，对于 vp+vp 的结构，在哪些情况下可以构成连谓关系，决定是否构成连谓关系主要看哪一个 vp，构成连谓关系时中心词落在哪一个 vp 上。这些都是我们很难回答而又不得不回答的问题。在机器翻译中我们遇到的这类问题简直不胜枚举。而这些问题仅仅靠机器翻译研究人员是很难回答的，只能依赖于汉语语言学界作更深入的研究。

2.2.2 句子分析

每一个英语句子都有唯一的一个限定形式的谓语中心成分，汉语句子则随意得多。简单的汉语句子可以只有一个词（独词句）或一个短语（如“一九四九年十月。北京。”），复杂的汉语句子可以是一个很长的段落。汉语文章的句子之间没有明确的界限。很多地方随意用句号和逗号都不算错。有些人甚至习惯于一逗到底，最后来一个句号，也不能说他就错了。这给汉语分析又增加了一个困难。现有的机器翻译系统一般都是以句子为单位进行翻译的。遇到一个特别长的汉语句子，如果作为一个整体来处理，往往带来巨大的时空开销（在允许回溯的情况下），而且这样做也并不合理，而分成几个小句来处理，又难以准确地断句。

2.3 汉语的语法层次问题

构成书面语言的最基本单位是文字（如字母和汉字），而由文字构成书面语言的过程并不是简单的堆砌，而是可以划分成很多个语法层次。一般而言，这些层次大致包括：字母（汉字）——词素（语素）——词语——短语——句子——句群——段落——篇章。语言的这些语法层次并不是人为规定的，而是语言中的客观存在。

在不同的自然语言中，虽然的语法层次是基本相同的，然而各种语法层次的特点和重要性却不尽相同。

语法层次的存在为自然语言的分析提供了方便。实际上，几乎所有的自然语言分析算法都是按照词语——短语——句子这几个层次的顺序进行的。可以想象，如果所有的语言都没有这种语法层次，而是像中国的古籍那样，由汉字（或字母）直接组成大篇的文章，那么计算机分析真是无从下手了。

我们认为，汉语分析的困难，在相当程度上是由于我们对于汉语语法层次的认识不够清楚，所采用的汉语分析算法与相应的汉语语法层次特点不相适应所造成的。

2.3.1 汉语语法层次的模糊性

在英语中，由于存在形态上的差别，这些语法层次是很容易区分的。而在汉语中，由于不存在形态上的明显差异以及汉语的书写习惯问题，语法层次的区分就存在一定的困难，而这种困难就导致了汉语语法层次的模糊性。

由于汉语的词语之间没有空格做间隔，产生了汉字层与词语层的模糊性。这种模糊性导致了汉语词法分析的困难，包括切词的困难和未登录词识别的困难。

汉语的词语层和短语层之间也存在模糊性。一个明显的特征就是离合词，离合词合的时候表现为词，而分的时候连同其插入部分义表现为短语。

汉语的短语层和句子层之间更没有明确的界限。这是因为汉语句子的构造原则与短语的构造原则基本一致，从短语到句子只是一种实现关系。汉语句子中没有明确的中心动词。

汉语的句子和句群之间同样没有明确的界限。很多情况下句号换成逗号并不算错。

从以上分析可以看出，汉语中虽然也存在语法层次，但汉语语法层次的划分并不象英语那么清晰，而是具有一定的模糊性，这种模糊性是导致汉语分析困难的一个重要原因。

2.3.2 汉语的词语层次与英语词语层次的不平行性

英语词语的判定非常简单。凡是在句子中以空格隔开的字母串都是词。汉语词语的判定却很复杂。在有关汉语切分的国家标准中，关于什么情况下该切，什么情况下不该切，有一套复杂的规定。在实际的汉英机器翻译系统的词典建设中，词语的选择也是一个难题，有些词选入词典中可以使翻译达到比较好的效果，但又会造成一些不必要的切分歧义。

英语词素组合成词的方式非常简单，词素与词素之间没有复杂的关系。汉字组合成汉语词的方式却非常复杂，有主谓、述宾、述补、定中、状中、联合等等。实际上，汉语词语、短语、句子的构成方式是基本相同的。

英语词语是不可拆分的，英语的词素都必须先组合成词语，然后才结合成短语。汉语词语却不一定。典型的情况就是离合词和语素字。离合词和语素字的情况说明，汉字可以

不经过词语层次，直接和其他汉字和词语组合成短语。

由此我们可以看出，汉语词语所处的语法层次与英语词语是不同的，汉语词语具备了很多英语短语才具有的特点。

2.3.3 汉语句子层次与英语句子层次的不平行性

在现有的形式语法体系中，句子作为一个独立的语法层次往往具有重要的意义。在Chomsky语法理论中，句子作为推导的起始符号，是定义一个文法的四元组中的一项。所有的语法分析算法也是以得到一个句子作为分析的结束。这种情况反映了句子层次在英语（以及其他一些语言）语法中的重要地位。

而在汉语中，句子并没有这么重要的地位。因为汉语句子的构造原则与短语的构造原则基本一致，从短语到句子只是一种实现关系，所以汉语的主谓结构实际上属于短语层的范畴，与英语中的句子没有对应关系。汉语中以句号（包括问号和叹号）结尾的“句子”实际上与英语的句子也不是处在同一个语法层次，而是在很多情况下对应英语中的一个或多个句子，即句群，有时也可以是单个的词或短语。

2.3.4 现行的汉语分析算法与汉语语法层次的特点不相适应

现有的自然语言分析算法所采用的分析层次都是：词语——短语——句子。相应的汉语分析算法也分为词法分析和句法分析两个阶段。在词法分析阶段，基本的分析算法都是基于乔姆斯基的三型语法，即正则语法。而在短语和句子层次采用的基本算法是乔姆斯基的二型语法，即上下文无关语法。

现行的汉语分析算法采用的也是这种模式。然而，由于汉语语法层次的特点，简单地套用这种模式并不合适，汉语语法分析的很多困难也是由此造成的。

英语的词法分析是非常简单的，而汉语的词法分析却要处理词语切分、未登录词识别、离合词、语素字等问题，非常复杂。在实际的汉英机器翻译系统中，词法分析程序的程序量甚至比句法分析程序还多得多。造成这种情况的原因就在于汉语词语层次与英语词语层次不平行。汉语词语的构成规律实际上与汉语短语的构成规律是基本一致的，用简单的正则语法无法处理，而我们现在采用的是打补丁的方法，为词语切分、未登录词识别、离合词和语素字的处理都编制专门的程序，结果是算法变得特别复杂，效果也不理想。

汉语句子层次与英语句子层次的不平行性，同样也给汉语的分析造成了问题。由于汉语的句子常常对应着英语中的句群，很多在英语中句子中不太常见的语言现象，在汉语的句子中变得很常见。典型的是汉语的承前省略现象，在汉语句子中非常常见，[8]一文对这种现象作了比较深入的分析。这种现象仅仅用上下文无关语法很难处理。

2.3.5 对汉语分析算法进行改进的初步设想

由于汉语的语法层次和英语的语法层次存在不平行性，而现有的汉语分析算法是根据英语的分析算法照搬过来的，很多地方不符合汉语语法层次的特点，因此我们认为，有必要对汉语的分析算法进行改进。

对于汉语的词语层次与英语的词语层次不平行的问题，我们认为改进的方法是：把汉字层次而不是词语层次作为句法分析的起点，把词语切分、未登录词识别、离合词和语素

字等问题都放到句法分析阶段，用统一的算法框架进行处理。这样做的原因是汉语词语的组成规律与汉语短语的组成规律是类似的，这种规律本来就是简单的正则语法所无法刻划的，而必须用上下文无关语法来刻划。

对于汉语的句子层次与英语的句子层次不平行的问题，我们认为应该在句子层次适当引入对汉语子句间关系的分析，如话题转移分析，成分省略分析等等。只有这样，才能对汉语的句子得到一个比较完整的认识。

以上只是我们为克服汉语分析中的困难提出的两个初步设想，具体实施的时候肯定会遇到各种各样的问题，我们将进行进一步的研究。

3 汉语到英语转换的难点

汉英机器翻译翻译中，汉语到英语的转换和英语的生成实际上是一个信息增加的过程。在汉语中所没有的各种形态信息，如单复数、时态、语态等等，在转换和生成中都必须添加上去。以下是在汉英机器翻译中遇到的最常见的几个比较难于解决的问题：

(1) 英语名词的单复数问题。在汉语句子中没有明确表示数量的词的情况下，我们一般都把汉语名词译成单数形式，这样做在很多情况下并不合适。

(2) 英语动词的时态问题。汉语没有时态，只有一些对时态有提示作用的助词（如“了”、“着”、“过”）和时间词（如“昨天”、“去年”、“将来”）、副词（如“已经”、“正在”）等。但大多数情况下，这些词语和英语时态之间并没有准确的对应关系。如汉语动词+“了”的结构有时应译成一般过去时，有时却应译成现在完成时或过去完成时，有时“了”又不表示任何时态，而这几种情况却几乎没有办法区分。对于没有这些词的汉语句子，则往往只能译为一般现在时，而这在很多情况下并不正确。

(3) 英语的语态问题。在英语中表示与事实相反的假设或个人主观愿望时，要使用虚拟语气。相应的汉语句子在翻译成英语时如果不采用虚拟语气，产生的英语译文就会与源文意思有很大的出入，甚至完全不知所云（如将“如果我是你”翻译成“If I am you”）。而汉语句子中并没有相应的语法标志，人只能根据句子所处的上下文环境和语义及常识进行判断，这对计算机来说是个非常困难的问题。

(4) 句子翻译问题。如前所述，汉语的句子层次与英语的句子层次是不平行的，汉语的一个句子译成英语时常常要分成多个句子。这种转换依赖于我们对汉语句子中的各个子句间的关系进行比较深入的分析。

(5) 英语冠词的添加问题。汉语没有冠词，在英语的名词短语中却往往要加上冠词，这其中又分为零冠词、定冠词和不定冠词三种情况。在翻译中要给名词短语加上合适的冠词也是非常困难的。

[9]中采用基于规则的错误驱动学习算法给英语译文添加冠词，取得了较好的效果。我们认为这种方法可以推广，用于解决汉英转换中的其他一些问题。如英语名词的单复数问题，汉语“了”字对应的英语时态问题等等。

4 结论

以上着重分析了我们在汉英机器翻译研究中遇到的一些难点，并对其中的一些问题表达了我们的看法。这些分析虽然比较零碎，不成体系，但也是我们经过一段时间思索的结果，希望能引起同行们的兴趣。

汉英机器翻译之所以滞后于英汉机器翻译的原因，一方面当然是由于汉语本身的特点（关键是没有形态变化）造成的，另一方面也是由于我们对汉语的认识还不够深入。我们衷心希望，汉语语言学界的学者们能和机器翻译界的研究人员能共同合作，把我们对汉语语法和对汉英机器翻译的研究都提高到一个新的水平。

解决汉英机器翻译所面临的困难从根本上还要依靠更加深入细致的研究工作，另外一种的办法就是采用受限汉语。由于受限汉语可以大大减少甚至消除汉语中某些类型的歧义现象，将使汉英机器翻译的正确率得到很大的提高，从而使汉英机器翻译基本上达到实用的水平。实际上，人们在很多场合下要求文章写得简洁明了、通俗易懂，这也是一种“受限”的情况，只是这种“受限”的要求没有形式化而已。如果能有一套比较合理并得到普遍接受的受限汉语规范，不仅可以提高机器翻译的正确率，而且对于人来说也可以减少很多不必要的误解和纠纷。这在一些不要求文学性而只要求文字清晰明了的领域，如商用信函、产品说明、规章制度、专利文献、标准文献等方面都有着广泛的应用价值。采用受限的方法可以在不改变现有算法的基础上较大地提高翻译的正确率及可读性，具有较好的使用价值。

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] 刘群，詹卫东，常宝宝，刘颖，一个汉英机器翻译系统的计算模型与语言模型，智能计算机接口与应用进展——第三届全国智能接口与智能应用学术会议论文集，电子工业出版社，1997

[3] 周强，现代汉语语料库多级处理与汉语短语结构分析，北京大学硕士论文，1993

[4] 周强，汉语语料库的短语自动划分和标注研究，北京大学博士论文，1996

[5] 俞士汶，朱学锋，王惠，张芸芸，《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

[6] 俞士汶，朱学锋，受限汉语研究的必要性，《语言现代化论丛》第三集，南开大学出版社，1997

[7] 白硕，论语义重心偏移，1998，待发表

[8] 宋柔，基于前缀省略的汉语叙述文篇章结构模型，第一届全国计算语言学联合学术会议，1991

[9] 常宝宝，刘颖，刘群，汉英机器翻译中的短词处理研究，中文信息学报，已录用，1998

作者简介：

刘群，副研究员，1966 年生，现就职于中国科学院计算技术研究所。研究兴趣为自然语言处理和机器翻译。当前研究方向为通用机器翻译开发平台和汉英机器翻译系统。

俞士汶，教授，1938 年生，现为北京大学计算语言学研究所副所长，研究兴趣和当前研究方向为计算语言学和机器翻译。

汉英机器翻译的难点分析1

引证文献(2条)

1. 吴云芳. 常宝宝. 詹卫东 汉英双语短语信息数据库的构建[期刊论文]-术语标准化与信息技术 2003(4)

2. 马红妹. 王挺. 陈火旺 汉语篇章时间短语的分析与时制验算[期刊论文]-计算机研究与发展 2002(10)