

英汉机器翻译软件长句分析刍议

刘志杰

北京创新未来科技有限公司

摘要：英汉机器翻译系统中有许多难点问题有待解决，其中英语长句一直是一个难题。如果我们能把英语长句切分为短句，然后再给机器翻译软件翻译，译文结果应该是较为理想的。本文简单分析英语长句偏大多数的原因，提出了切分英语长句子的方法，在实验中结果是很理想的。

关键字：机器翻译、英语长句、切分

Some Opinions on Analyzing Long Sentences for English-Chinese Translation Software

Liu Zhijie

Beijing Creative Next Technology Company

Abstract: In MT software, the super long English sentences are so much difficult for it to handle. Generally speaking, if the software analyzes the long sentences incorrectly, the result of translation must be wrong. In some machine translation software, they can deal with the short simple sentences successfully except the long sentences. Therefore if the super long sentences can be divided into some short simple sentences, the machine translation software will manage them quite easy. And the result of translation maybe is very good. In this article, the author puts forward some basic methods to cut the long sentences into some short ones. This will do good to MT software. At first, this method can improve the analyzing ability for the software. Second, it can assure the quality of the result of translation.

Key words: machine translation software, long sentences, segmentation

一、概述

机器翻译软件商品化已经有十余年的历史了，但是同其它应用软件相比，它还没有形成大气候，虽然有的专家说机器翻译软件是信息高速公路上的加油站，但还不是人们必备的工具软件。相反，一个电子词典软件人们对它的评价却很高，确实给用户解决了不少的语言障碍问题，也很有市场。原因很简单，机器翻译软件译文质量、译文的易懂性还不尽人意。机器翻译技术中有很多难点还没有突破，其中之一是长句的分析是比较大的难点。把长句分析错了，译文肯定是错误的。现在市场上的几种翻译软件，对短句的翻译都是可以的，一旦遇见了长句，表现都不是很好。所以，我们有必要在翻译前做个长句的预处理工作，分析一下英语的长句，搞清楚英汉两种语言的最主要差异，用一些拆分的方法将英语的长句分成简单句，再进入机器翻译软件进行翻译，这样译文结果应该是很理想的。

二、英语长句产生的一些主要原因及切分的必要性

做任何语言对的翻译软件，我们都得细致研究这两种语言的特点，搞清楚外语长句产生的一些内在因素，是为了让机器很好地分析；摸清汉语的特点，是为了让机器在分析完长句以后做汉语生成。那么，说到拆分英语长句，我们应该明白英语长句偏多的原因是多方面的，但是，最主要的原因有以下几方面：

第一：英语词语之间的语法关系，除了通过安排词序来表达外，还常常采用各种各样表示关系和连接的手段，如介词（词组）、连词、关系代词、关系副词、连接代词、连接副词、非谓动词（词组）、词的形态变化（如词尾变化、格的变化）等等。

第二：另一方面，能充当英语句子成分的，有长短几乎不加限制的短语和从句，而且从句中还有从句，再加上各种并列成分，附加成分（如同位语、插入语、独立成分），尤其是形形色色的修饰成分，通过丰富而灵活的连接手段，使英语的句子更加复杂。

第三：英语句子呈句首封闭、句尾开放。定语修饰语可以后置，又有关系词与被修饰语连接，句子可以不断向句尾扩展、延伸。

第四：英语还常用先行代词“IT”及其它制定性代词，把真正的主语或宾语移到后面，并根据需要加以不断扩充，形成句子的末断重量。

第五：英语以主谓核心协调控制全句结构，有词的形态变化约束，有连接手段连接词语和从句，有代词使词语前呼后应。

以上是英语长句偏多的主要因素，这些因素都可以使冗长的句子不致流散，结构紧密。这也充分体现了英语是一种形合式的语言，靠各种连接手段把英语句字编织成网，环环相扣，疏而不露。但是，我们还可以得知，尽管英语句子长而又长，我们分析以后都可以把它们分为下列五种最基本的句型：

1. 主语+动词	The class begins. The baby cried.
2. 主语+动词+表语	We are teachers. She seems happy.
3. 主语+动词+宾语	He cheated me. The news surprised us.
4. 主语+动词+间宾+直宾	She dreamed a beautiful dream. I gave him a book.
5. 主语+动词+宾语+宾补	I painted the wall yellow. We considered him right.

英语各种长短句子，一般都可以看成是这五种基本句型及其变化、扩展、组合、省略或倒装。下面有个英语的长句，我们用人工翻译，机器翻译软件整句翻译，再用机器翻译软件分句翻译，看一下译文结果的不同所在。

It was our view that the United States could be effective in both the tasks outlined by the President --that is, of ending hostilities as well as of making a contribution to a permanent peace in the Middle East - if we conducted ourselves so that we could remain in permanent contact with all of these elements in the equation.

1. 人工翻译参考译文：如果我们采取行动以便能够与中东问题各方面始终保持接触，那么我们美国就能有效地担当起总统所提出的两项任务，那就是在中东结束敌对行动以及对该地区的永久和平作出贡献。这就是我们的观点。

2. 机器翻译软件的全句译文：美国能在由结束战斗以及在中东制造对永久的和平的贡献总统即提出的两个任务中是有效的，这是我们的观点- 如果我们表现了，我们才能在平衡方面保持与所有这些因素永久的联系。

3. 切分后机器翻译软件的译文：

(1) It was our view . 它(这)是我们的观点。

(2) The United States could be effective in both the tasks outlined by the President . 美国能在由总统提出的两个任务中是有效的。

(3) That is, of ending hostilities as well as of making a contribution to a permanent peace in the Middle East . 那是，关于结束战斗以及关于在中东制造对永久的和平的贡献。

(4) If we conducted ourselves . 如果我们表现了。

(5) So that we could remain in permanent contact with all of these elements in the equation . 我们才能在平衡方面保持与所有这些因素永久的联系。

以上的机器翻译结果我们可以得知，整句的翻译的结果叫人比较难以理解，易懂性比较差。而把原来的长句切分成五个短句后的译文结果看起来尽管零乱，但是我们也能从中获得这样的信息：我们的观点是：担当总统提出的两个任务→结束中东对抗→保持联系。

因此我们可以说原句要表达的最主要信息我们基本上搞懂了，机器翻译软件没有把这五个分句的译文结果合在一起组成一句汉语，但是其易懂性又基本上让人满意，这是由于汉语的特点所决定的。同英语相比，汉语更注重内在的意念而不重外在的形式，也就是说汉语是意合性的语言。所谓意合，是指词语与分句之间不用语言形式手段连接，句中的语法意义和逻辑关系通过词语或分句的意义表达。从前文分析英语长句多的原因我们可以得知，英语是一种形合性的语言。所谓形合是指句中的词语或分句之间用语言形式手段（如关联词）连接起来，表达语法意义和逻辑关系，所以说英语的句子更注重形式，注意结构完整，注重以形显义。这样，我们就得把英语中用形合法组成的句子拆分成简单的形式，而后在进入机器翻译软件进行翻译。

三、建立词典库和语法规则库，为切分长句做必要准备

将英语中的动词和功能词放入待建造的词典库中，对它们进行简单的描述，为编写分析规则做准备。

第一部分、建立常见词类的词典库

这个词典库不同于用于机器翻译过程中的词典库，它仅仅是为分析切分长句服务。它其中的词可以不带汉义，但是词类必需标注，语法/句法作用必需体现，兼类判断必需明细。

(一) 遵循英语的五个基本句型的原则，建立英语动词词典库

从英语的五个基本句型可以看出，英语动词在句中的作用是很重要的，它处于核心地位，为此要首先建立一部动词词典库，信息应该包括动词的“时、体、态、式”以及“性、数、格”的变化。同时还要考虑到动词本身的兼类问题，在建立动词词库时要有比较详细的信息来描述兼类的现象和解决的办法。解决的办法就是要建立一些语法规则，来分析判断它的

准确词类。还要把介词、副词同动词搭配后形成的动词词组放入动词词典库中，为以后的进一步分析打下基础。

(二) 建立名词词典库

英语的五个基本句型离不开名词的参与，名词是这五个基本句型中的很关键的语法成分。为切分长句而建立的名词词典库关键的地方是兼类判断。英语中名词兼类的现象同动词一样很普遍，因此，在构建名词词典库时对兼类的要做重点处理，尤其是那些能转变为动词的名词，要写一些判断规则来分析。

(三) 功能词是关键，建立比较详细的功能词词典库

前文我们分析了英语长句偏多的一些主要原因，也清楚了英语是一种用形合法组成的语言。如果要切分好长句，分析功能词是很关键的，那么功能词词典就显得尤其重要。英语中下列此类应该收入功能词词典。

1. 关系词和连接词：关系词包括关系代词、关系副词、连接代词和连接副词，如 who, whom, whose, that, which, what, when, where, why, how 等，它们功能是用来连接主语和宾语从句或表语从句。连接词包括并列连词和从属连词，如 and, or, but, yet, so, however, as well as, neither.....or, either.....or 以及 when, while, as, since, until, unless, so.....that, so that 等，这些词用来连接词、词组、分句或状语从句。英语的句子构成几乎离不开这些关系词和连接词。

关系词和连接词是英语中使用频率很高的两种词类，是机器翻译面对的难题，对这些词处理得是否合适，会直接影响到长句的分析与判断。一个英汉翻译系统处理长句子的能力如何，首先还是看这个系统是如何处理关系词和连接词的问题的。

这类问题解决起来更为复杂，这里只提一个基本原则：解决这类问题行之有效的方法必须考虑到这些词的共性的一面，又要考虑到它们的个性用法。举例说 WHAT 这个词在复合句中往往是身兼二职，如它可以引导宾语从句，在从句中又作宾语，为了分析的单一性，可采用复制技术，再复制一个到两个 WHAT，使每个都成为功能上的单义词，这样分析出来的结构关系就显得很清楚，切分才有道理。此外，解决这类问题要考虑这些词与动词的关系以及与语义的关系，尤其是要考虑这些词同动词的关系，同时要结合系统的语言分析规则，这样才能比较好地解决它们在句子的语法功能，才能把长句子拆开。

2. 介词：英语介词约为二百八十六个，分为简单介词（如 with, to, in, of, about, between, through）、合成介词（如 inside, onto, upon, within, without, throughout）和复合介词（如 according to, along with, apart from, with regard to）。介词是英语里最活跃的词类之一，是连接词、词语或从句的重要手段。可以这样讲，英语句型之所以纷繁复杂，是因为这些介词的用法十分灵活，有人把英语称作是介词的语言就是因为这一点。在基本词典中要充分考虑这一点。

一般说来，英语介词同一些实义词有着千丝万缕的联系，如英语中的 # V + # P 结构。所以在设计基本词典时要本着这样三个基本原则：

（第一）如果这个介词跟动词有固定的搭配关系，则要把这个介词放到具体的动词身上，不要把这类介词孤立起来去另外处理，否则不利于介词的处理与分析，随着介词用法的增加，基本词典也不断增大，不利于句子的分析判断。

（第二）要结合介词本身的用法，加上必要的语义分析，判断出介词的语法成分及句法功能。比如介词短语在句子中是做状语还是做定语，这类问题要考虑到语法分析和语义分析。例如在 NP1+PP+NP2 的结构中，我们先是依据 NP1 和 NP2 的语义分析解决定语分析的问题，然后再考虑它作状语的情况，这是为判断句型，做切分长句的服务的。

(第三) 建立介词分析的算法时必须考虑筛选(加权)和回溯, 例如表示支配关系的要比表示附加关系的加权级要高, 换句话说, 在分析过程中宾补可冲掉已得成份的定语, 倒装宾语的分析又必须靠动词的回溯分析来解决。所以说如果第一次分析时没有确定介词短语的语法成分及句法功能, 则要进行第二次分析, 直到这个介词短语分析成功为止, 这样才有把握来确定长句中的介词的功能, 才有把握找到长句子里包含的基本句型, 才有把握把长句切开。

(四) 建立标点符号词典库

标点符号(如逗号、引号、感叹号)是英语长句中的重要成分, 句中的插入语、非限定性成分都是由它们构成的, 所以标点符号词典库也是重要的环节, 是切分长句的一个标志。

第二部分、建立语法规则库

语法规则库中, 结合建立的词典库来构造。主要内容要涵盖英语中的五个最基本的句型。特殊的句型, 如 **There be** 句型、**It** 引导的句型不包括在这五个基本句型中, 做单独处理。建立语法规则库的基本理论依据是短语结构语法, 再结合一定的语义分析, 只要能 **liu101** 把句子切分正确, 其余的分析就交给机器翻译系统去做了。在语法规则库中主要分为以下几个模块:

动词模块 **名词模块** **介词模块** **WH 词模块** **连词模块**

模块之间的关系不是对立的, 而是相互联系, 互为依存的。尤其在做名动/动名的兼类判别时, 模块之间是能相互调用的, 这样才不会使语法规则库无限度的增长, 有了量的控制, 才有质的保证。还有, 动词的语法规则应该充分考虑到其形态变化及语法结构的变化, 比如: **I make money** 和 **Money was made**。这两个句子的分析结果应该是一样的, 至于它们怎么生成汉语, 是下一步机器翻译系统分析的问题, 在此语法规则库能判断他们是动词, 结合前后的关系, 把长句分开就算完成任务。

语法规则库最大的用处是判断当前分析的句子是否是复合/复杂句。我们以句子结束的标志符号为依据, 凡是一个句子中有两个或两个以上的连接词或关系词, 那么这种句子就有切分的必要, 这种句子是复合句/复杂句。

总之, 在做词典库和语法规则库时, 肯定要涉及到对英语词汇的分类。分类的基本原则是宜粗不宜细, 太详细了, 参数过多, 不利于两个库的构造与日后维护。建库最根本的目的只是做句型判断, 而不是做更进一步的分析。

四、如何切分英语的长句

总的原则是: 结合已经建立的词典库, 对一个英语长句首先从句子后边开始扫描做预处理; 而后再从句首扫描, 完成子句的分析; 最后切分出短句。

(一) 从句子尾扫描

本次任务是确定句中的名词、动词、功能词以及逗号, 根据词典中提供的信息确定它们语法作用, 结合语法规则库, 语法成分能充分确定的, 则把分析的信息保留下来; 不明确的, 留作下一步分析。

(二) 从句子首扫描

第一步确定名词、动词、功能词和逗号的语法功能后，把查到的信息带入语法规则库，从句首开始分析，根据语法规则库提供的英语五个基本句型，把长句子切分出来。具体原则如下：

1. 遇到功能词的（介词和连接词除外，下文再述。）句子有两种情况，一种是整个长句中
没有逗号，另一种是整个长句中有逗号。遇到没有逗号的长句时以此功能词为中心
向此功能词后去找出主语+谓语+（宾语）的结构，如果主谓宾结构中有附加成分，比
如主语或宾语的定语，谓语的修饰语等，这些附加成分都要算作这个主谓宾结构中；
如果遇到功能词前有逗号，以逗号为基准，把这个功能词引导的句子切分出来，再归
入英语的基本句型中；如果功能词前有逗号，功能词引导的句子后面还有逗号，则逗
号间的句子算作一个基本句子，结合句中的动词，分析后归纳到基本句型当中等待处
理。从句中省略引导词的，要进入语法分析规则，将省略的引导词通过分析复制或补
译出来，再做切分（分析方法同 1）。复制或补译完全是分析上的需要，也是形式上的
需要。不这么做，我们没有办法来明确语法关系，当然就谈不上切分了。
2. 遇到介词，要做两种分析。如果这个介词同动词形成了短语，则要把它划给动词，作
谓语使用；如果不是这样，是其它词的附加成分，那么就把它划给其它词类。比如说
划给句中的名词，此时要判断介词或介词短语做什么成分，从而确认了名词的语法功
能。
3. 遇到连接词如 and, or 时，要通过语法规则库判断它们是否连接两个或两个以上的成分
（包括词典库中的动词），如果是的话，把它们连接的成分归于一个分句中处理。如
果它们连接的是两个或两个以上的从句，那么，以引导从句的引导词为界限，把这个
从句拆分成子句。
4. 遇到逗号的情形有两种：
第一：如果句中只有一个逗号，那么就以此逗号为根据，把这个长句子切开，然后再
根据英语的基本句型将切开后的句子归类；
第二：如果一个长句中有两个或两个以上的逗号，这时要做如下处理：一是逗号与逗
号之间的句子如果没有动词，则按短语处理，按原样保留；二是逗号与逗号
如果有动词，则要进入语法规则进行基本句型的分析，最后输出分析结果。

(三) 一些特殊的句型，暂时不能归入英语的基本句型的，进入特殊句型库分析处理，然后
输出切分结果。

(四) 简单的复合句不做切分，直接输出结果，送给机器翻译软件处理。原因是简单的复合
句对翻译软件来讲，要容易处理一些，机器翻译软件中建立这类句子的规则，分析这
类句子相对来说容易一些。简单复合句是指整句中没有附加成分，没有插入语，更没
有复杂的并列结构。如：What you see is what you get.

(五) 下面是一个切分的例子：

原长句：As the patient gets older, as the coronary arteries become narrowed, and particularly
if the blood pressure continues to rise, the day will come when the myocardium is no longer
equal to the strain, and congestive heart failure develops.

针对这个长句子，第一次从句子尾开始向句子首查，分别查出以下重要信息：

Develops动词	become.....动词;
Is.....系动词	as.....功能词引导从句;
When.....功能词引导从句	gets.....动词;
If.....功能词引导从句	gets.....动词;

在查出这些关键词的基本信息以后，进入语法规则库，再从句首开始向句尾分析，将分析的结果输出：

- (1) As the patient gets older.
- (2) as the coronary arteries become narrowed.
- (3) and particularly if the blood pressure continues to rise.
- (4) the day will come when the myocardium is no longer equal to the strain.
- (5) and congestive heart failure develops.

最后，把切分后的分句送入机器翻译软件进行翻译。

笔者始终认为，英汉机器翻译软件翻译出来的句子之所以可懂性较差，主要是因为受到英语长句子的困扰。机器翻译不可能什么句子都能处理，花很大力气去分析长句子，还不如把长句子切分成基本句子再去翻译。这样做的结果会不会让人满意，我想如果用户是一位以汉语为母语的人，他是会理解机器翻译给他的译文的。用汉语的人都知道，汉语句子偏短，表达灵活多样，层次也比较分明，如果把英文的长句子给机器翻译来翻译，译文综合又欠妥，给用户一大段不带标点符合的汉语，其可懂性一定是很差的。在汉语里，我们可以说“我吃饭了”，也可以说“饭我吃了”，这是因为汉语是意合性的语言，字里行间我们会悟出它要表达的意思。英语这种语言长句很多，我们人工在翻译长句时也是要重新组织其顺序或结构，以汉语的习惯表达出来，我们一定不会把长句子翻译成从头到尾连个标点符号都没有汉语。人工翻译是这样，那我们就更没有理由去难为机器翻译了，结合英汉两种各自的特点，该分的分，该合的合，那我认为还是很值得的。

五、小结

以上提出的几种方法只是一种尝试，但是笔者认为，对一个英汉机器翻译软件来讲，最难对付的恐怕就是英语中的长句子。这些长句子最好还是译前切分一下再进入机器翻译软件中处理，否则整句放进去，搞出的译文逻辑很乱，可懂性比较差，有些根本就看不懂。译前做这些长句的切分，可谓是划整为零。目前，做机器翻译软件的方法大致有几种：一是以语法分析为主，语义分析为辅；二是从语料库的角度去做；三是从语义的角度去做；四是用模板与插槽的办法。但是无论哪种办法实现的翻译软件，对短句的翻译都还可以，即使有些短句翻译错了，可能是因为语义分析、兼类判断、歧义结构等方面欠妥，但大部分软件对短句的处理能力都是很强的。所以，笔者建议机器翻译软件的开发者，要处理好英语中的长句，还是把它切成短句为好，仅仅依靠有限的语言规则去对付不胜枚举的长句，恐怕是行不通的。

参考资料

1. 连淑能，英汉对比研究。北京：高等教育出版社，1993。
2. 张培基，李宗杰等，英语翻译教程。上海：上海外语教育出版社，1983。
3. 刘学功，英语词语分离。上海：上海交通大学出版设，1991。
4. 黄凯，科技英语结构与翻译模式。广州：华南理工大学出版社，1992。
5. 论英汉翻译技巧论文集。北京：中国对外翻译出版公司，1986。