关于机器翻译系统未来的方向

田中康仁
日本兵库大学
E-mail:yasuhito@humans-kc.hyogo-dai.ac.jp

摘要： 从实用的机器翻译系统开发开始，时间大约已过去 20 年了。在此期间，有许多机器翻译系统问世。本文想就今后的机器翻译系统的方向进行一些讨论。

关键词： 机器翻译、自然语言处理、评价

Future Direction of Machine Translation System Development

Yasuhito Tanaka
Hyogo University in Japan
E-mail:yasuhito@humans-kc.hyogo-dai.ac.jp

ABSTRACT: Twenty years have passed since full-fledged machine translation (MT) system development began, and numerous MT systems have been developed in the intervening years. This paper focuses on the future, and discusses likely future directions of Mt system development.

Keyword: Machine Translation System, Natural Language Processing, Evaluation

1. 机器翻译系统的现状

在日本大约有 20 多家公司在开发机器翻译系统，其产品数量大约有 50 个左右。大多数是英译日、日译英的机译系统，也有几家公司在做日译韩、韩译日、日译汉、汉译日的机译系统，再有少数几家公司在做英语、韩国语、汉语以外的语言的机译系统，及开发以英语作为中间语的机译系统。即使是开发时间较长的英译日机译系统，其翻译的正确率也就是在 70% 左右。

机器翻译系统的评价：

用日本电子辞书研究所开发的 EDR 语料(英文)，对两词～七词句的数据进行评价，得到以下的结果：

	5 分	4 分	3 分	2 分	1 分	总分数	平均值
两词句	15	1	1	1	1	19	4.47
三词句	336	97	16	11	0	460	4.65
四词句	1,369	336	157	25	2	1,889	4.61
五词句	3,655	808	510	53	4	5,030	4.60
六词句	4,742	1,379	595	81	1	6,798	5.58
七词句	4,471	2,331	870	118	1	7,791	4.43
合　计	14,588	4,952	2,149	289	9	21,987	4.53
%	66.35	22.52	9.77	1.31	0.04	100%	

上表中，其分值高的表示机译的质量好。纵观这些情况则还有以下两个问题：

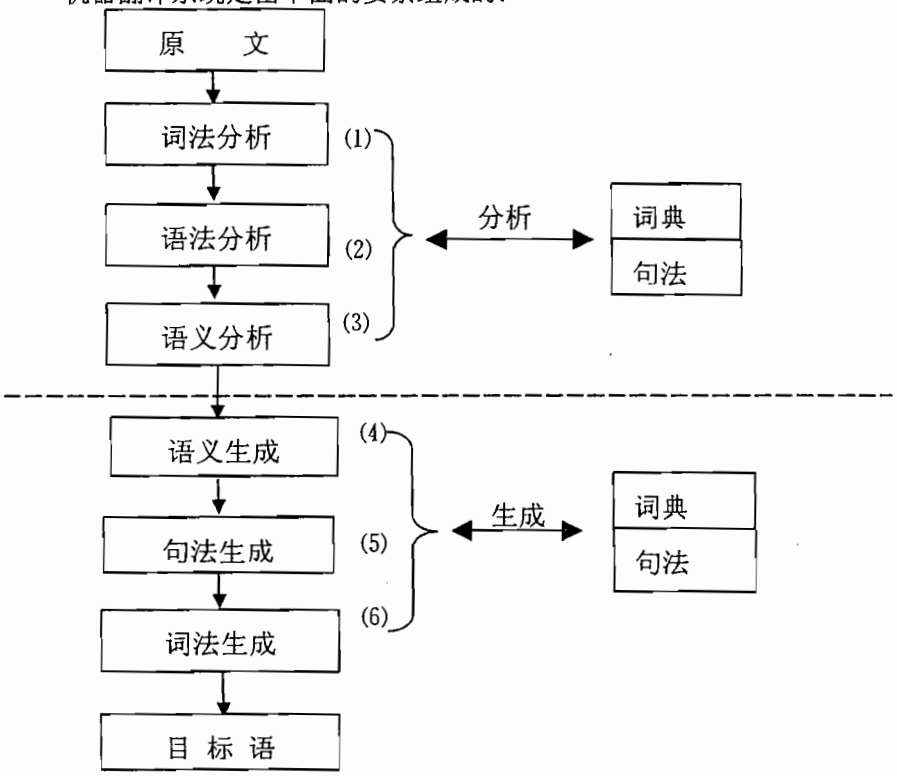
① 怎样提高已开发的机器翻译系统的精度呢？

② 如何把还未研究的语言对应起来呢？

2. 机器翻译系统的问题及提高精度的意义

(1) 机器翻译结构上的问题

- 机器翻译系统是由下面的要素组成的：



大致由六个元素组成，假设这 6 个元素的可信度分别记作 P_1、P_2、P_3、P_4、P_5、P_6，则原文到目标语的正确转换的可信度就是：

$$P=P_1 \cdot P_2 \cdot P_3 \cdot P_4 \cdot P_5 \cdot P_6$$

这里以 1.中的评价结果为例，有

$$0.7=P_1 \cdot P_2 \cdot P_3 \cdot P_4 \cdot P_5 \cdot P_6$$

因 P_1、P_2、P_3、P_4、P_5、P_6的概率不能个别测定，所以假定这 6 个元素的可信度是相同的，

$$P_1=P_2=P_3=P_4=P_5=P_6$$

那么，　$0.7=P^6$

$$P_1=\sqrt[6]{0.7} \approx 0.942286$$

也就是说，可信度平均值是 0.94，用百分比表示即为 9 4 %。由此看出要想提高系统整体的可信度，就必须提高每个元素的可信度。但是，实际上各个要素的概率值并不相同，所以开发者最好要知道系统中哪个要素的概率值低，这样有针对性地提高那些低概率的要素的可信度，就会提高整体的效果。姑且不考虑开发费用和时间，改良的最优先顺序是可以判断的。

◆ 关于翻译系统各构成要素

各构成元素是由词构成的。因日语、汉语、泰语等是由词构成的黏着语，所以要有词语的切分操作。假设各词的可信度用 P_{ij} 表示，其中 i 是各构成要素，j 是在该构成要素中的第 j 个词的要素。如果假设各要素的可信度为 0.94，则有 $0.94=\prod_{j=1} P_{ij}$，所以也还必须提高各单词的可信度。

(2) 机器翻译系统的问题及提高精度具体几个内容

如何提高可信度?

(2-1) 考虑单词级

i）充实词典，减少未定义词

为此有必要对大量的数据和语料进行分析，收集新词。分析报纸、WWW 主页及各专业杂志、语料。

ii）收集复合词和专业术语

一定要把构成复合词和专业术语的词结合成一个整体后生成译词

日语	英语
限界利益	marginal gain
限界容量	limiting capacity
価格限界	price limit
対象限界	bound of object
能力限界	limit of ability
規格限界	bound of standard

因此，要积极地增加复合词和专业术语。对于日语来说，许多日语学者把复合词或术语切分为更小的单位（词）。由于机读词典的容量不受限制，所以应该多加一些固定词组，这是解决歧义的一种办法。采用长词后，构成句子的词数也就减少了。

(2-2) 收集短语

英语句子中有动词短语、副词短语、介词短语、名词短语。短语是一个语义的集合。收集短语也可以减少句子的构成元素。

例如： [介] at home 家

[名] hope of succeeding 成功の見込み

但是也不能说收集长词就一定好，必须注意到这一点。日语中长词常用做名词，也有当成サ变动词使用的。举出英语中一个著名的例子：

① It rains cats and dogs.

② I have cats and dogs.

“cats and dogs”与 “rain” 共现时和与 “have”共现时的意思不同，这种使用上的差别一定要在词典中作出标记。

(2-3) 收集并列短语

收集用 and 和 or 的并列短语。在英语中用 and 和 or 连接的，译成日语时，译词的词序会有不同的时候。

A and B → B A 、 A B

A or B → B又はA 、 A又はB

例如：ladyies gentlemen → 紳士淑女

另外，A and B C 的场合 → AC and BC
→ (A) and (BC)

大部分翻译系统是采取用规则的方法进行处理，今后采取登录到词典的方法以减少歧义。

(2-4) 句法

ｉ）句法分析

把句子分成单句、复句、重句。除此之外还可分成陈述句、疑问句、感叹句。调查英语的句法树与日语的句法树是怎样对应的。调查构成句子的每个动词也是重要的。说起句法树，也必须调查是单词层次上的句法树好呢，还是句子层次上的句法树好？因此，去除了两个语言间的句法上的歧义的树库是必要的。而且，规模要大。

制作两种语言的大量句法树库，并且将这些树按照句子的种类和动词进行分类，对语言间的变换的方法进行探讨。几乎还没有发现能简单完成这样的一系列作业的研究。今后，收集两个语言间的句法树库就是为了构建“语言变换的语法”。

ii）配价语法数据的收集

应该大量地收集每个动词的价文法，这是一种语义分析。

例：[A] provide [B] for (to) [C] →[A]が[C]に[B]を与ぇる。

[A]:[C]是人 [B]：是物

iii）表明动词级别

日语和英语的动词的数量都很大。日语、英语都有特定的动词和名词搭配，还有动词化。

日语： 结婚⊕する → 结婚する

英语： make a sign → サインする

工业用英语句子与一般的英语句子在使用动词方面有区别，依照动词的重要性应该给出相应的标记 A,B,C。其中对于 A,B 级动词要调查各类词典，应该找出它们的用法、格语法，还应该找出它们的习惯用法。另外就是从英语与日语对译的的语料库中调查。这样若干人系统地调查，也要花上几年的的时间。

对于 C 类动词也应如此照办，这些都是需要费用和人力。

(2-5) 句子

在机器翻译系统中有把原文句子和译文句子一一对应起来（即双语对齐语料库），根据句子的排序输出译文的实用化系统。这样的系统的优点是省去了复杂的句法分析，因此提出下列的方案：

ｉ）多语言语料库的制作

应该准备庞大的多语言或并行的对齐语料库。这些语料库一定要包含的各种各样领域的语料。例：

good morning → おはようございます

How do you do? → ごきげんいかがですか

没有必要对这样的句子进行分析。在英语中应该收集不算 a，the 的 1 0 个词以内的惯用句。

ii）特殊句型或特殊句型的抽取

在某个特定的领域（如：新闻和天气预报）里，收集使用特定的表现形式，在翻译系统中巧妙地使用它们，就可以较容易地生成质量好的译文。

用离散形 n 元语法这种简单的方法，可以抽出特定句型。如把数字部分抽象成[Num]也是一种方法。例：

There are 2 books. → There are [Num] books.

除此之外，把 地名 → [Loc] 时间→[Time]

人名 → [PN] 日期→[YDM]

当在一个句子里有两个以上的人名时，可用[PN-1]、 [PN-2]……表示，还可用[PN-m]、[PN-f]来区分人的性别。

iii) 句法和用词分析用的多连接语料库

最好是多语言的语料，但应该收集大量的并行语料，来解决形态分析、组合分析，消除歧义，最好再加上一般句和工业用句。

iv) 使用同类词变换句子

有的句子是可以用其他类似于该句的部分词做替换，还有的找出类似的句子修改部分词。

He is a [teacher]. → 彼は[先生]である。

替换：He is a student. → 彼は学生である。

为了这个目的就需要收集大量的并行语料，所以必须考虑使用并行语料。至此我们总结所讲述的各种情况，得到：

i) 知识数据性能的抽取 ii) 扩充词典

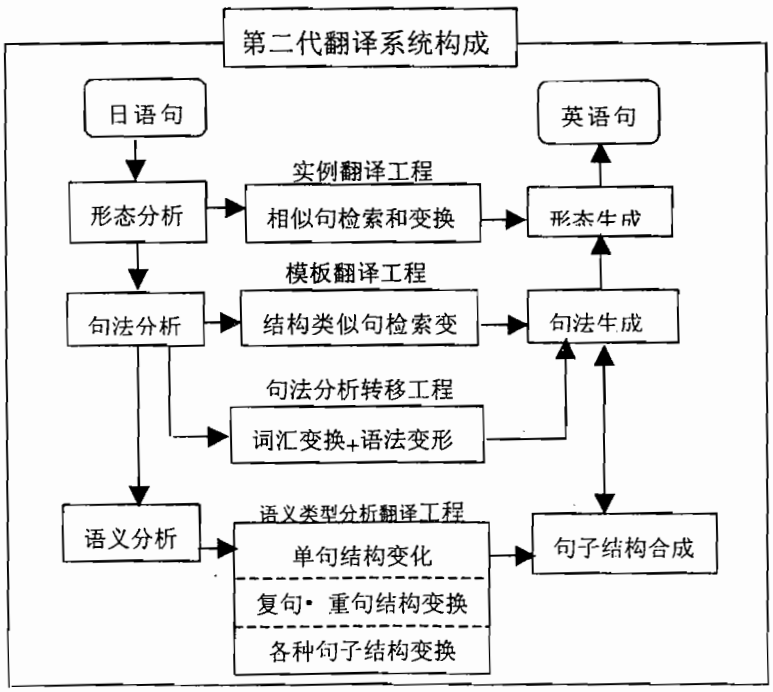
iii) 句法的大系化 iv) 充实各种大量的多语言语料库

这些问题也是今后连续的大课题。

(3) 对机器翻译系统进行全面的再考察

目前，我们在以语法分析为核心的机器翻译系统中，增加语料库，把我们头脑中的这些东西一体化。

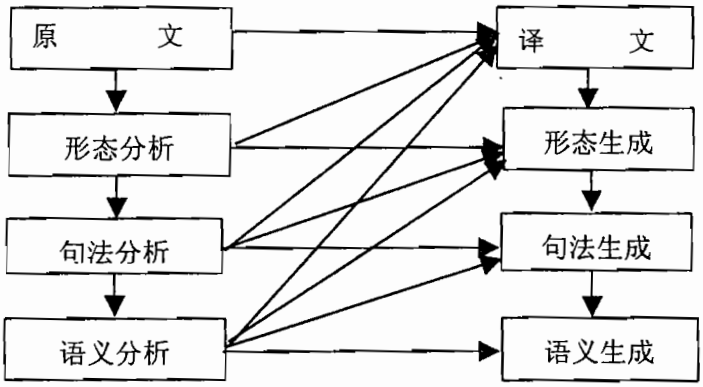
鳥取大学的池原教授提出了下面的混合型方法。



这也只是一个方法。作者可以考虑更加特殊的情形，例如最好是由日语直接生成英语，把日语句子作为 key 直接查出相应的英语句子、输出地道的英语句子。还有就是边进行日语的语义分析，边进行英语的形态生成。

最好能把考虑到的各种方法的可能性，编制成实际的程序、制成表格和文件。在我们头脑中池原先生提出的那种新的系统，应该是进行更柔性的处理。

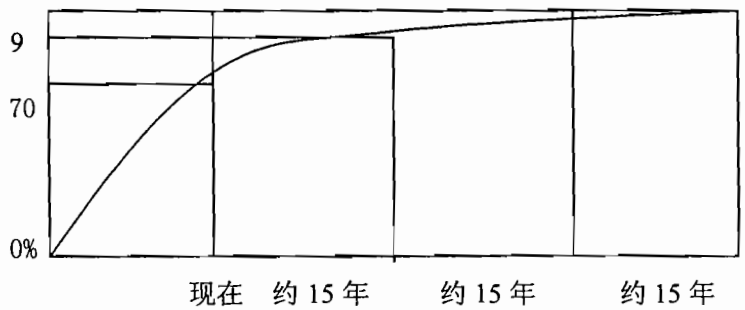
作者推荐的翻译系统的构成：



这份原稿登出来并投稿到某个国际会议，同时出席了 1999 年 11 月 1 日～3 日，在中国北京举行的全国第五届计算语言学联合学术会议（JSCL～99），会议论文集中的论文画出的图。这个内容没有发表。

在这篇论文之前的德语→汉语、俄语→汉语机器翻译的研究者们也是向图中那样考虑的。

(4)　机器翻译系统的质量像下面的曲线画的那样慢慢地提高



今后 10～15 年间，如果努力开发，扩充机器翻译用的知识库，就会达到 90%左右。再过 10～15 年就可以达到 95 %或更高一些。如果把翻译系统限制在某个领域里，说不定会更快地实现这一目标。

最近，图书的电子化工作急速地发展，对电子版本进行再编辑应用到词典机器翻译用专门用语词典简单制作

今后的个人计算机速度更快，如果是这样处理的时间就会缩短，可以很容易地处理更大量的数据和知识数据，这样具有提高机器翻译系统的精度吧。

如果文本数据一下子扩大 2 0 倍，也许原来大约需要 1 5 年的时间就可以缩短。同时如果词典数据也在一年里投入过去的 2 0 倍，那么翻译的质量就会提高。

由于机器翻译系统的精度年年有改进，所以要想得到一定量的错误数据，还必须增加比今天更多的文本数据，所以调查问题也变得困难起来。

现在机器翻译系统可以说精度在７０％左右，一般说来这比人工翻译的速度提高了３倍。因而提高机器翻译精度，会有什么变化呢？

所谓70%的精度也就是说剩余的30%必须由人手工地加以修改，理解比人工翻译提高３倍的效率的意义。我们把翻译的精度设为 q，其中：0≤q≤1。那么，1/(1-q)就表示比人工翻译提高了几倍的效率：

精度　70% → 提高 3 倍
　　80% → 提高 4 倍
　　90% → 提高 10 倍
　　95% → 提高 25 倍

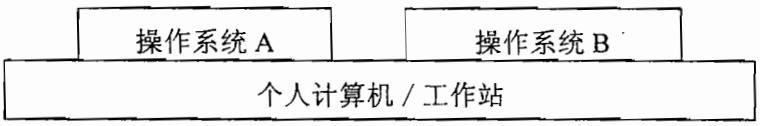
这里只是单纯地表示出来，从准备各种适用的工具和词典，到要达到上面的精度，还不知道提高效率最快时间是几时，希望很快能提高到接近９５％。

3．如何处理已开发的机器翻译系统以外的语言

对于已开发的机器翻译系统，我们要考虑处理除英语、汉语、韩国语以外的语言，例如阿拉伯语、波兰语、蒙古语如何处理好呢？考虑哪种语言，必须考虑那种语言在多大程度上与日语有关系这样一种社会意义。

例举出首先必须考虑的项目。

ｉ）操作系统



我们每天使用的个人电脑或工作站，但是国家不同其操作系统也就不同，即使同样的Windows，在日本进行可日语版。使用 A 国语言的人也将 Windows'98 做了相应的修改，使用日文版的 Windows'98 必须能处理 A 国语言。

ii）输入方式、键盘、文字字体

操作系统在什么环境下使用是个大问题，下面的问题是：如何输入 A 国语言的文字。

① 输入 A 国文字的方式用 A 国标准确定，那么它与日文的输入可以共存吗？

② 为了输入 A 国文字，定出键盘标准，使得即能输入 A 国文字，也能用于输入日文。

③ A 国文字的显示是否正常？

iii）编辑

日语及 A 国文字在屏幕上显示是否正常？有处理日文和 A 国语言的标准的编辑器吗？大多数的文字是从左向右书写，但阿拉伯语是由右向左书写，要转换这样的语言时，编辑器也要做很大的改动。

iv）词典的制作

要处理 A 国语言，首先必须要制作机读词典。还需要 A 国语言的词典和日语与该词典的对译词典。同时词典中要有基本词汇和分领域的专业词汇。

ｖ）语料的制作

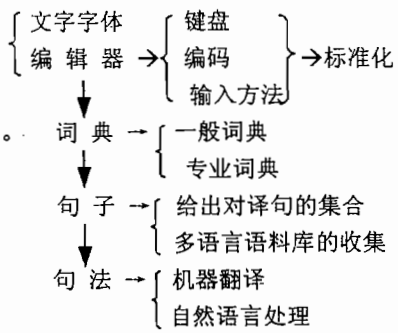
与词典一样，必须收集制作句子集。不仅给出该词单独的使用，还要有在句中作为的构成要素的使用方法。因此，有必要给出实例告诉该词如何使用？起什么作用？另外不仅给出单纯一个语言，还有必要给出对译的句子集。

vi）句法

按照 A 国语言进行自然语言的处理，就必须把 A 国语言的句法符号化、体系化，还必须检查句法的正确性，这还有很多工作要做。

用这些工具，杂配套的机器上就可以处理语言了。

解决以上各个方面就可制成右侧的流程图：



4. 今后，我们前进的方向

我们既要提高以日语、英语为中心的机器翻译的精度，也要进行处理其他语言的基础研究、实验、模型制作，必须推进这两方面。世界上有各种各样的语言、民族，只使用日语和英语是不可能的。

早稻田大学有 ICMTP(International Conference on Multilingual Text Processing)学会，请参考。

URL:http://www.mling.waseda.ad.jp/icmtp/

E-mail: icmtp@mling.waseda.ac.jp

5. 结束语

由于 E-mail 的发展，我们的世界显得变小了。然而通信上的语言是五花八门，开发一种便于大家使用的语言转化工具是一个重要的想法。在这里我们期待着开发出这样一种工具。

参考文献

[1]（社）日本电子工业振兴协会："关于自然语言处理系统的动向的调查报告"，平成 9 年 4 月。

[2] 牧野武则："评价技术 《机器翻译》Bit 别册"，共立出版，1988 年 9 月。

[3] 长尾真："机器翻译到什么样就是可能的？"，岩波出版，1986 年 6 月。

[4] Language and Machines: Computer in Translation and Linguistics. National Academy of Sciences-National Research Council (1996)

[5] Annual Report of Ikehara Laboratory 1998，Natural Language Processing, Vol. 3，Tottori University, Faculty of Engineering (Ikehara Laboratory)

[6] 傅爱萍："英汉机器翻译的转换生成策略"，计算语言学文集（JACL-99 Proceeding），清华大学出版社，1999 年 10 月。

[7] 邸海燕、柴佩琪、许玉祥："基于汉语语料库和规则库的德汉复合句的转换生成"，计算语言学发展与应用，清华大学出版社，1995 。

[8] 李向东、周清波等："智能型俄汉机器翻译系统的句法规则库的设计原理"，中文信息学报，1999。

注：数据

英文数据用到的是日本电子辞书（株）的英文语料

关于本论文翻译的基础：有关易理解性的详细内容参考参考文献（3）P55

其原文在 Language and Machines: Computer in Translation and Linguistics. National Academy of Sciences-National Research Council (1996) (ALPAC 报告)中。