机器翻译常见错误类型总结[1]

--- CWMT2013 机器翻译人工评测结果再分析

赵红梅　刘群

中国科学院智能信息处理重点实验室　中科院计算所　北京　100190

E-mail：{zhaohongmei, liuqun}@ict.ac.cn

摘 要：本文在 CWMT2013 人工评测的基础上，对人工评测中有关错误类型的评测结果进行了总体分析，重点、细致地分析和统计了英汉机器翻译最常见的几种错误类型以及产生这些错误的原因等，同时深入分析和比较了基于规则的系统和基于统计的系统在这几种错误类型上的表现差异，旨在更全面、深刻地揭示机器翻译目前存在的主要问题，进而为机器翻译系统质量的提升提供较为全面、客观、细致的数据参考。

关键字：机器翻译、常见错误类型、译文选词错误、实词译文缺失、词序不对、译文和原文意思相反

Common Error Analysis of Machine Translation Output

Hongmei Zhao and Qun Liu

Key Lab of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS, Beijing 100190, China

E-mail：{zhaohongmei, liuqun}@ict.ac.cn

Abstract: *Based on the manual evaluation for the 9th China Workshop on Machine Translation, this paper gives a further detailed analysis and statistics on several common error types appearing in the generated translations, and discloses the main sources of these errors. Meanwhile, this paper also presents the performance differences on each common error type between rule-based MT system and statistical MT system.*

Keywords: *Machine translation, common error type, incorrect words, missing content words, wrong word order, translation with the meaning contrary to the original.*

1 引言

第九届全国机器翻译研讨会（CWMT2013）机器翻译评测在英汉和维汉新闻机器翻译项目上设置了人工评测。人工评测采用由参评单位参与的“众包”的方式。每个评测结果分别由三位评测者从忠实度和流利度两个方面进行打分。同时还要求评测者对所打分的句子进行错误分析，即对以下选项进行单选或多选：1）译文选词错误；2）实词译文缺失；3）词序不对；4）译文和原文意思相反；5）命名实体问题；6）数量词/时间词问题；7）其它错误； 8）没有错误。最终评测结果表明：在两个项目上，前三种错误类型最为常见。关于 CWMT2013 人工评测的具体内容请参考《CWMT2013 机器翻译评测报告》[赵红梅等，2013]。

本文作者在以上人工评测的基础上，分别对基于统计和基于规则的英汉机器翻译系统的译文进行分析。重点抽样考察了被判断有上述前四种错误的机器翻译译文。针对这些译文的错误展开了进一步的分类和统计，揭示了这些错误发生的原因，以期能更全面、细致地了解机器翻译的现存问题，从而为业界同行改善机器翻译系统质量提供有意义的数据参考。

2 错误类型的确立及定义

作者在确立错误类型方案之前，特意对不同翻译方法/模型的英汉机器翻译译文进行了实际的错误分析和统

[1]**基金项目：**国家自然科学基金（61379086）

计，并参考以往研究工作的知识和经验、英汉和维汉机器翻译的具体特点以及研究者们比较关心的错误类型等，制定了英汉和维汉人工评测错误类型分类方案。方案出来后经过几位专家的认真讨论而最终得以确立。

对于七种错误类型以及“没有错误”的判断，我们分别定义如下：

1) **译文选词错误：**指原文单词有对应译文，但译文错误或者不准确；

2) **实词译文缺失：**指缺少应有的实词译文，主要包括以下几种情形：1’ 译文消失：指实词原文缺少必须翻译出来的相应译文；2’ 原文照搬：实词译文为不符合翻译要求的原文照搬；3’ 省略未译：没有根据翻译需要翻译出原文中省略的实词。

3) **词序不对：**指原文中有的单词或单词串的意思虽然翻译出来了，但在译文中出现的位置不正确，导致译文不流畅或令人费解；

4) **译文与原文意思相反：**指译文有的单词或单词串的含义与原文完全相反，这是在信息传递中最不能让受众接受的错误之一；

5) **命名实体问题：**指原文中的命名实体（包括人名、地名、机构名、组织名等）翻译有错误或没有翻译；

6) **数量词/时间词问题：**指原文中的数量词或时间词翻译有错误或没有翻译；

7) **其它错误：**我们把译文中出现的不属于上面任何一类的错误均归为“其它错误”，在评测打分工具中，鼓励评测者对此类错误进行简单的书面描述。

以上七种错误中，前三种类型是比较概括的类型。实际上，一般翻译错误主要属于这三大类型（“出现多余译文”的错误很少，故我们没有加进来。虚词缺失的现象一般也不明显，所以我们只重点关注了“实词译文缺失”）。其后的三种错误类型则主要是研究者们一直比较关注和重视的错误类型。除了“译文与原文意思相反”以及“数量词/时间词问题”之外，我们的错误类型分类以及最常见的几种类型在评测提交译文中出现的错误比例与国际上其他同行的机器翻译错误分类[David et al., 2006; Ariadna et al., 2005]及评测结果[David et al., 2006]是基本一致的。在人工评测的实际操作中，我们允许评测者在对译文打分时采用多选的方式，这是因为一个句子可能包含多种错误，而且前六种错误类型并不是相互独立的，前三种和后三种之间有可能存在交叉，比如“命名实体问题”就有可能同时属于“实词译文缺失”或者“译文选词错误”或者“词序不对”。

8) **没有错误：**指译文与原文意思完全吻合，译文表达完整，挑不出错误。我们在对人工评测结果进行统计时人为规定：只有当三个评测者对某一句话都判断为“没有错误”时，才记之为“没有错误”。

3. 错误类型评测结果的粗分析

在对评测结果进行分析之前，我们先定义以下四个概念：

1） 词错率

词错率 = 实错词数 / 总词数

其中，实错词数是指一段原文中发生某类翻译错误的单词个数，总词数是指这段原文的单词总数。

2） 句错率

句错率 = 实错句数 / 语料规模

其中，实错句数是指一段语料中实际发生某类翻译错误的句子数量，语料规模是指这段语料的句子总数。

CWMT2013 机器翻译评测英汉和维汉的人工评测语料规模为 500 句（内容分别选自 CWMT2011 的英汉新闻测试语料和 CWMT2013 的维汉新闻测试语料）。每个句子都由三名评测者分别进行打分。这样对一个系统来说，每类错误出现的最大频次为 500*3 次，而每类错误的实错句数最多为 500 句（语料规模）。

3）判断一致性

句错率的大小除了受语料本身的影响，还受评测者判断一致性的影响，如果不同评测者判断分歧大、判断一致性低的话，句错率可能会产生虚高的现象。而某类错误出现的频次与其实错句数的比例越高，说明评测者对该错误类型判断的一致性越高（此次评测中最高值为 3）。我们将这个比例进行归一化处理（即除以每句被判断的次数，此次评测为 3），得到的比值便可作为评测者对错误类型判断一致性的衡量指标（简称为“判断一致性”）：

判断一致性 = 错误频次 / （实错句数 * 判断次数）

4）误判率

人工评测存在一定的主观性，也可能会出现一些判断失误的情况，在重新检查人工打分情况时，我们将被误判为某一错误类型的句数与被判断为该类错误的总句数的比值，作为衡量该错误类型判断准确性的一个指标：

误判率 = 被误判为某类错误的句数 / 被判为该类错误的总句数

本小节错误类型分析中的句错率数据是建立在 CWMT2013 机器翻译人工评测基础上的一种粗分析数据。“粗”的原因是：1) CWMT2013 人工评测是以句而不是以词为单位，所以涉及到词的错误被放大到句子一级了，比如“译文选词错误”、“实词译文缺失”等，显然结果被严重放大了；2）人工打分的误差因素没有考虑进来。所以本小节涉及的句错率数据仅具有相对意义，更深入细致的分析数据请参考第 4 节的内容。

3.1 不同项目错误类型的表现差异

此次人工评测各错误类型的句错率及判断一致性如表一。

表一 CWMT2013 机器翻译评测英汉和维汉项目各错误类型的句错率及判断一致性

判断类型	英汉				维汉			
	出现频次	实错句数	句错率	判断一致性	出现频次	实错句数	句错率	判断一致性
译文选词错误	9212	4056	0.90	0.76	8081	3839	0.85	0.70
实词译文缺失	6436	3297	0.73	0.65	8536	4006	0.89	0.71
词序不对	6094	3158	0.70	0.64	3008	2136	0.47	0.47
译文与原文意思相反	744	570	0.13	0.44	248	205	0.05	0.40
命名实体问题	2005	1343	0.30	0.50	1540	1000	0.22	0.51
数量词/时间词问题	468	319	0.07	0.49	677	490	0.11	0.46
其它错误	1040	936	0.21	0.37	148	141	0.03	0.35
没有错误	177	59	0.01	1	342	114	0.03	1

注：“没有错误”类的“实错句数”及“句错率”应为“实际句数”与“出现比例”。

从表一可以看出：

1) 在英汉和维汉两个项目上，“译文选词错误”、“实词译文缺失”和“词序不对”均为最常见的错误类型；

2) 三个最常见的错误类型中，除了维汉项目的“词序不对”之外，评测者的判断一致性均高于或接近于 0.66，这意味着在这三个错误类型上，三位评测者中差不多有两人意见是一致的，也就是说这样的错误类型判断是基本可靠的；

3) 维汉新闻“词序不对”的句错率大大低于英汉新闻。通过后面的分析以及请教从事机器翻译的维族老师，我们发现：英文中存在位于中心词后面的修饰词（定语和状语），而汉语译文要求修饰词位于中心词之前，这二者词序不一致是导致英汉“词序不对”句错率偏高的主要原因。而维语中基本不存在这个问题（维文中修饰成分一般位于中心词前面，与汉语一致）。另外，尽管维文是主宾谓（SOV）结构，与汉语的主谓宾（SVO）结构不同，但是借助于统计机器翻译的语言模型，能够较好地解决维汉两种语言主谓宾结构不一致的问题，所以相对来说，维汉翻译中“词序不对”的错误较少；

4) 维汉新闻“实词译文缺失”的句错率明显高于英汉新闻，这应该与两个因素有关：1’ 英汉新闻项目评测组织方提供的训练语料规模为 500 万句对，而维汉新闻项目的训练语料规模仅为 11 万句对，二者相差悬殊；2’ 维语的词语形态变化比英语更为复杂丰富。

3.2 不同翻译方法/模型错误类型的表现差异

3.2.1 不同翻译方法/模型的总体表现差异

表二中给出了每个项目人工评测得分最高的前五个系统。其中，忠实度和流利度的评测打分采取五分制，具体打分标准和计算办法请参考《CWMT2013 机器翻译评测报告》[赵红梅等，2013]。

从表二中可以看出：

1）基于规则的系统在英汉新闻项目的人工评测忠实度指标上具有一定的优势；

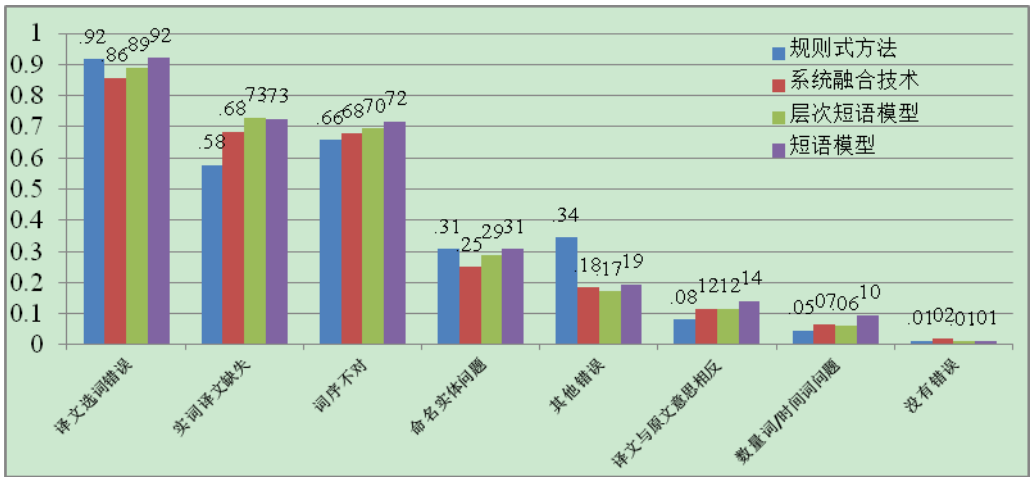
2）基于统计的系统中，采用系统融合方法的系统（系统融合系统）得分高于层次短语系统，层次短语系统得分高于短语系统。

表二 CWMT2013 不同机器翻译方法/模型的人工评测结果（前五）

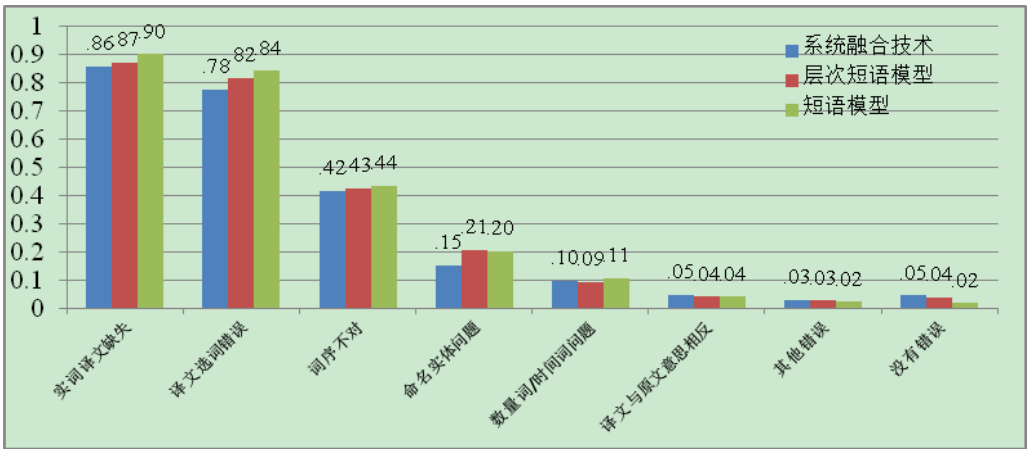
英汉新闻			*维汉新闻*		
系统	*忠实度*	*流利度*	*系统*	*忠实度*	*流利度*
EC 规则系统	3.26	3.01	UC 系统融合系统	3.42	3.11
EC 系统融合系统	3.11	2.97	UC 层次短语系统 A	3.30	3.03
EC 层次短语系统 A	3.00	2.86	UC 层次短语系统 B	3.12	2.83
EC 层次短语系统 B	2.98	2.80	UC 短语系统 A	3.12	2.82
EC 短语系统	2.93	2.76	UC 短语系统 B	3.10	2.77

3.2.2 不同翻译方法/模型在错误类型上的差异

我们仍以表二中列出的每个项目的前五个系统为例，图二和三给出了每种不同的方法/模型在每类错误上的句错率（如果同一项目上两个系统采用的是同一种翻译模型，则给出的是二者句错率的平均值），从图二中可以看出：英汉项目中规则式机器翻译系统“实词译文缺失”的比例明显低于统计式机器翻译系统，“其它错误”的比例则大大高于统计式机器翻译系统，在“译文选词错误”和“命名实体问题”上规则式系统不具备优势。



图一 英汉机器翻译不同翻译方法/模型句错率的对比图



图二 维汉机器翻译不同翻译方法/模型句错率的对比图

4 英汉新闻项目四种错误类型的进一步分析

4.1 译文选词错误

4.1.1 同一段语料不同系统的“译文选词错误”对比

作者任意选取了英汉新闻项目测试集中一段语料对采用不同方法的机器翻译译文进行考察，此段语料（简称为 A 语料）原文共有 51 句（含 1224 个单词），规则系统和系统融合系统在此段语料上分别有 50 句（含 1205 个单词）和 47 句（含 1109 个单词）有“译文选词错误”，表三和四给出了针对这两个系统这些语句存在的“译文选词错误”的具体分类统计。

作者根据分析的具体情况，将所有的“译文选词错误”按产生的原因分为以下几类：

1）**消歧错误：**此类错误是指词本身存在歧义，机器翻译在歧义词的译文选择时出现了错误。如：

原文：The soldiers took bin Laden's **body** with them.

机器译文：(这些)士兵带上本·拉登的**[身体;遗体]**.

此例中，“body”有“身体”、“尸体”和“遗体”等多种译文，机器应当选择“尸体”。

上例为此次参评的规则系统的译文，该系统的消歧策略之一是：当很难解决词的歧义问题时，系统就采用中括号的方式将多个译文用分号并列呈现。不过，作者在进行译文错误分析时只参考中括号内的第一个译文。

2）**兼类错误：**此类错误是指原文单词的兼类现象造成的译文错误。如：

原文：Environmental scientist Chris Bowser pulled a tiny shrimp-like creature from the muck in an eel **trap** as teenagers in chest waders surrounded him in the rushing Fall Kill, where they were collecting transparent baby eels.

机器译文：环境学家克理斯加油车作为青少年的[箱子;胸部]拉一种微小类似小虾的来自一条鳝鱼**[困住;引诱]**的凑钱的生物[涉水者;高统靴]秋季在慌乱中包围他引起死亡,他们[(在)哪里;那里](正在)收集透明幼小的鳝鱼.

参考译文：环境科学家克里斯 •鲍泽将一条小虾一样的生物从鳗鱼**陷阱**的淤泥中拉出来,而穿着及胸防水裤的几位十几岁的青少年正围着他站在名为 Fall Kill 的湍急的溪流中,他们正在这里收集透明的鳗鱼幼崽。

此例中的 trap 应该翻译成它的名词形式（译文为“陷阱”），但误译为其动词形式“[困住;引诱]”。

3）**固定搭配错误：**此类错误是指原文中存在的固定搭配所造成的译文翻译错误。错误表现在：有的是将固定搭配按单个单词拆开进行翻译，有的是固定搭配本身有歧义导致翻译错误，还有的是不应该理解成固定搭配却被当作固定搭配来翻译，等等。如：

原文：If she lives **as long as** her mother she could go another 15 years, 16 years, which will put Charles around about 78, Dickie Arbiter, former press secretary to the Queen, told the Royal Diary.

机器译文：“如果她能去生活,**只要**她母亲的另一个 15 年, 16 年来,将“查尔斯大约有 78 迪基自己的仲裁者、前新闻秘书告诉皇家日记。

as long as 在此句中不应当作固定搭配。

原文：On Friday, President Obama gave the order for the U.S. military to carry out the mission to **take out** Osama bin Laden.

机器译文：上周五,奥巴马总统命令,美国军方执行本·拉登**拿出来**的任务。

take out 在此句中应该译为“除掉”。

由于固定搭配涉及到几个单词，作者考察了部分固定搭配的翻译情况，发现组成固定搭配的单个单词一般翻译都不正确，所以我们在统计固定搭配的“出错词数”时按它们所涉及的单词总数来计算。

表三 规则式系统在 A 语料上“译文选词错误”的分类统计

错误类型	词错率	具体分类	出错词数	词错率	举例
消歧错误	0.064	名词	39	0.032	protocals, body, scales
		动词	15	0.012	see, follow, started, promoting
		命名实体	9	0.007	Bowser， Reyes-Bravo，Pastor，ICC
		形容词	6	0.005	most, prospective, common, accepted

		副词（WH 词）	4	0.003	what, when, where, whatever
		副词（普通）	1	0.001	globally
		介词	4	0.003	off, as, with, to
		连词	1	0.001	so
固定搭配错误	0.012		15		set their minds, take out，one-day runs（6 个）
兼类错误	0.003		4		trap
总	0.079				
误判句数	9	误判率	9/50=18%		

表四 系统融合系统在 A 语料上“译文选词错误”的分类统计

错误类型	词错率	具体分类	出错词数	词错率	举例
消歧错误	0.059	名词	34	0.028	capital, awareness, compound, match
		动词	20	0.016	started, banned, practicing, engaging
		命名实体	7	0.006	人名:5 国名:1 缩写:1
		形容词	4	0.003	relevant, rushing, common
		副词（WH 词）	2	0.002	who, where
		副词（普通）	1	0.001	after
		介词	1	0.001	for
		连词	2	0.002	after
固定搭配错误	0.016		19		brings together, as long as, take out（8 个）
兼类错误	0.001		1		vowed
总	0.076				
误判句数	9	误判率	9/47=19%		

4.1.2 不同语料同一系统的“译文选词错误”对比

为了研究不同的语料对“译文选词错误”造成的影响，作者在系统融合系统的翻译结果中选择了另外一段语料（原文 60 句，共 1218 个单词，简称 B 语料），系统融合系统在此段语料中有 50 句（含 1089 个单词）出现“译文选词错误”。表五给出了系统融合系统在这段语料上“译文选词错误”的分析情况。

表五 系统融合系统在 B 语料上“译文选词错误”的分类统计

错误类型	词错率	具体分类	出错词数	词错率	举例
消歧错误	0.037	名词	17	0.014	American, British, anyone
		动词	14	0.011	Dominated，betting，crumble
		命名实体	2	0.002	Discover，Ridgeline
		形容词	9	0.007	deepest, anal, thinning
		副词（WH 词）	1	0.001	what
		副词（普通）	1	0.001	secondarily
		介词	1	0.001	by
		连词	0	0.000	
固定搭配错误	0.019		23		let…slide, knock off, given way to（10 个词组）
兼类错误	0.003		4		what, as, perfect, barrel
总	0.059				
误判句数	8	误判率	8/50=16%		

各种不同原因造成的不同语料、不同翻译系统的“译文选词错误”的词错率还可以参考图三。

从表三、表四、表五和图三可以看出：

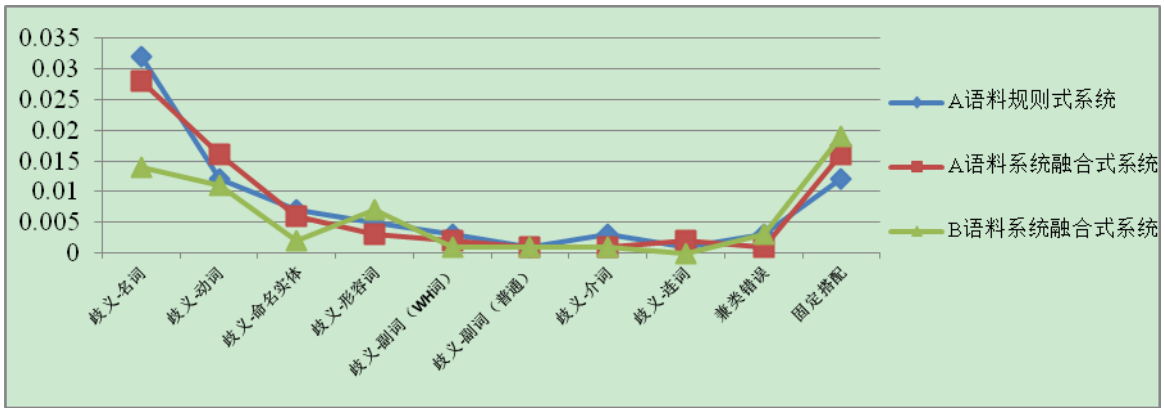
1）“译文选词错误”的词错率随语料的不同而发生变化，在以上两段语料中为 5%～8%左右；语料不同，不同原因造成的错误分布也不相同，语料对错误原因分布的影响比翻译方法的影响更大。

2）在“译文选词错误”产生的原因中，词的歧义占大多数，其次是固定搭配；

3）词的歧义错误主要来源于名词和动词这两大类，名词歧义造成的错误比动词歧义更多；

4）相对于统计系统，规则系统的“消歧错误”和“兼类错误”更多，而“固定搭配错误”较少（这应该与规则系统自带较成熟的知识库有关），规则系统在“译文选词错误”上不具备优势；

5）此类错误误判率为 16%～19%。



图三 各种不同原因造成的不同语料、不同翻译系统的“译文选词错误”的词错率

4.2 实词译文缺失

作者任意选择了英汉测试集原文的一段连续语料（共 75 句，1502 个单词，简称语料 C）来考察不同机器翻译系统的“实词译文缺失”错误。在这段语料中，规则系统和统计式系统（仍以系统融合系统为例）分别有 48 句（共 1105 个单词）和 50 句（共 1092 个单词）被判断有“实词译文缺失”的错误。我们仍然以有译文缺失的原文实词的单词数与这段语料的单词总数（即 1502，为方便计算，作者在此处没有采用这段语料中实词的总数，这样计算的“实词译文缺失”的词错率比实际的的词错率略低）的比值作为“实词译文缺失”的词错率。

我们考察的“实词译文缺失”错误主要有以下几类：

1）**译文消失**：单词有原文，无译文，且翻译要求译文必须出现；

2）**原文照搬**：单词有原文，有原文形式的译文，但不符合一般翻译要求。该类又分为命名实体（简称“原文照搬-NE”）和普通词（简称“原文照搬-OD”)两 个小类；

3）**省略未译**：单词原文省略，无译文，但翻译要求译文翻译出原文中省略的词。

表六 规则式系统在语料 C 上“实词译文缺失”错误的分类统计

判断情况	词错率	具体分类	缺失词数	词错率	举例及说明 （第一行为原文，第二行为机器译文，第三行括号内容为错误说明）
正确判断	0.014	原文照搬-NE	17	0.011	原文：Are **Beyoncé** and Co. to blame? **Beyonc•é** 那帮家伙应负责任吗?. . （Beyoncé 碧昂丝，有此类错误的词包括人名 14 个，地名 2 个，车名 1 个）
		原文照搬-OD	3	0.002	The "us"-centric songs of the 1980s, like "Ebony and Ivory," have given way to "me"-centric lyrics like Beyonce's "It's **blazin**', you watch me in amazement." 20 世纪 80 年代以“我们”的歌曲,像“黑檀木和象牙,有“我”的方式——像碧昂丝是“以民为本的歌词 **blazin** ”,你吃惊地看着我。 ”
		省略未译	1	0.001	Tornadoes devastate South, killing at least **290** 龙卷风摧毁南方,杀害至少 **290 个**. （译文缺量词“人”）

误判		译文选词错误	15	0.010	**Brain** types booze more -- are you surprised? **大脑**类型狂饮更多--你感到吃惊(吗)?. .
		可不译	7	0.005	In addition to Sunday's television broadcast , Logan will also appear in an interview on 60 Minutes' webcast, **60MinutesOvertime.com**. 除了星期天的电视广播以外,洛根也将在 60 分钟的网络广播上出现在一次[面试;采访],**60MinutesOvertime.com**. （有此类错误的词包括公司的网站名 6 个，如 Facebook，链接 1 个）
		词序不对	5	0.003	She reported without incident for nearly an hour before her interpreter heard words in the Arabic-speaking crowd that gave him pause. 她在她的翻译之前没有遇到麻烦报告近一个小时在讲阿拉伯的人群(其使他停下来想一想)里听见言语.
误判句数	27	误判率	27/48=56%		

表七 系统融合系统在语料 C 上“实词译文缺失”错误的分类统计

判断情况	词错率	具体分类	缺失词数	词错率	举例及说明 （第一行为原文，第二行为机器译文，第三行括号内容为错误说明）
正确判断	0.018	译文消失	9	0.006	The decision is **made** in a short story in "Action Comics" No. 900. 这一决定是在很短的故事在“漫画行动” 900 号。 （原文 made 的译文消失）
		原文照搬-NE	13	0.009	Last year, he launched the **Thiel** fellowship, which gives grants as large as $100,000 to 20 tech entrepreneurs who drop out of college by age 20 to pursue their own ideas. 去年,他发起了 **Thiel** 奖学金,资助 10 万至 20 万美元大科技企业家大学辍学的 20 岁前去追求自己的想法。 （有此类错误的词包括 11 个人名，2 个地名）
		原文照搬-OD	4	0.003	And there was no lack of **self-lovin'** in the '80s. 在 80 年代,就不缺少 **self-lovin** 。 （ self-lovin 未译）
误判		译文选词错误	12	0.008	Why do smart kids grow up to be **heavier** drinkers? 为什么聪明的孩子长大成为**较重**的饮酒者?
		可不译	6	0.004	**Facebook** may be bigger, but it "doesn't have the focus on coupons or local business relationships" to knock off **Groupon** or **LivingSocial**. **Facebook** 可能会更大,但并没有把焦点放在券或本地业务关系”要减掉 **Groupon** 或 **LivingSocial** 。 （有此类错误的词包括 6 个公司网站名，如 Facebook）
		词序不对	5	0.003	Lara Logan breaks silence on Cairo assault 拉拉劳根打破沉默的开罗攻击
误判句数	23	误判率	23/50=46%		

通过表六和表七我们可以发现：在本段语料中，

1）“实词译文缺失”的词错率为 1.4%～1.8%；

2）“实词译文缺失”类错误中，**命名实体**没有得到正确翻译的占较大比重，这个现象在规则系统中更为明

显（词错率为 1.1%）；

3） 统计系统中“译文消失”的词错率较高（0.6%），而规则系统没有发现此类错误；

4） “实词译文缺失”类错误误判率较高，为 46%～56%。

4.3 词序不对

作者任意抽取了英汉测试集中的一段语料（共 83 句，含 1750 个单词，简称语料 D）考察“词序不对”错误。在这段语料中，规则系统和统计式系统（仍以系统融合系统为例）分别有 47 句（1199 个单词）和 50 句（1282 个单词）被判断为“词序不对”，作者通过考察发现这两个系统被误判的句数分别为 8 句和 7 句(即实际出错句数分别为 39 和 43 句)，由此计算出误判率分别为 17%和 14%。由于“词序不对”可能涉及到句子的多个部分，不方便计算词错率，所以只能计算规则系统和统计系统在这段语料上“词序不对”的句错率，分别为 47.0%和 51.8%。通过对这些错句的逐个分析，作者发现规则系统和统计式系统分别有 66 处和 69 处“词序不对”错误（有些句子存在多处“词序不对”的错误）。表八和九中的比例为各分类在“词序不对”错误中所占的比值。

在错误分析过程中，我们借鉴了 David Vilar 等人的做法[David et al., 2006]，为每个词序错误区分了两个维度：1）**错误长度：**发生词序错误的长度，这个长度只取两个值：单词或者词串；2）**错误距离：**发生词序错误的词离正确位置的距离，这段距离只取两个值：长距离或者短距离。长短距离没有绝对的界限，作者的定义是：如果词序只是在同一个简单的语法成分中发生错位（比如：在没有较复杂修饰关系的名词词组内部），就称之为短距离词序错误，反之就是长距离词序错误。具体的统计结果请见表十。

另外，状语除了语法分类外，还增加了语义分类（区分了时间、地点和其它）。

作者在分析“词序不对”的错误时，具体分到哪一子类，主要依据：1）词序错误产生的原因；2）发生词序错误的词/词串在句中所担任的语法成分或者结构。大部分情况下这两个依据是一致的，小部分情况下二者不一致，这时作者一般将词序发生错误的原因作为分类的依据，如下例：

原文：The way you design and maintain your home could **play a role** in whether you pack on the pounds **or** keep them off.

机器译文：你设计并且[维持;坚持]你的家(庭)的生活方式能在你是否储存(这些)(英)镑中**起着作用**或避开他们.

上例中发生词序错误的部分是“play a role”，但是发生错误的原因是“or”导致的结构歧义，所以我们将这个词序错误归为“其它”错误中的“and/or”类。

表八 规则系统在语料 D 上“词序不对”错误的分类统计

分类	频次	比例	具体分类	频次	举例及说明（第一行为原文，第二行为机器译文，第三行括号内容为错误说明）
状语	25	0.38	介词结构	15	If your kids seem to be saying "I" and "me" **with excessive frequency** these days, you may want to check their iPods. 如果你的小孩似乎说我和我**[带着;用]过度频率**目前,你可能想检查他们的 iPod.
			状语从句	4	who soon became separated from her team and bodyguard **as the crowd swept her up**. (其[马上;不久]与她的团队和保镖失散)**随着人群席卷她**.
			其它	6	Logan's story will be broadcast on "60 Minutes" **this Sunday, May 1 at 7 p.m. ET/PT**. 洛根的故事将在 60 分钟被广播**这星期天,5 月 1 日下午 7 点 ET/PT**.
			时间	10	略
			地点	2	略
			其它	13	略
定语	18	0.27	定语从句	7	But at least he's "challenging the cultural assumptions **that cause a lot of people to make bad life decisions**." 但至少他(正在)挑战文化假设**(致使许多人(们)做糟糕的生活决定的)**. .
			介词结构	3	Instead of focusing on items **like discounted clothing**,

					[与其;非但不]**像对服装打折扣一样**[集中在;关注][项目;条款;产品],
			其它	8	Plus, Deals will create "real-world, non-virtual" uses for Facebook's online currency **called Credits**. 另外,交易将建立“真实世界,非虚拟”的用途的 Facebook 的在线货币**叫做信贷**。
其它	23	0.35	命名实体词组	7	says Satoshi Kanazawaat at **Psychology Today**. 在**心理学**上 Satoshi·Kanazawaat **今天**说. (此类错误包含人名 4 个，杂志名 2 个，会议名 1 个）
			标点	4	Plus, Deals will create "**real-world, non-virtual**" uses for Facebook's online currency called Credits. 并且,交易将建立**现实世界,**因为 Facebook 网站(脸谱网)的在线货币**非虚拟的**使用叫[信任;荣誉;契约]. (此处是 real-world 后的逗号引起的结构歧义。这类错误包括逗号、破折号带来的结构歧义以及成对引号的利用问题）
			and/or	3	but it "doesn't have the focus on **coupons or local business relationships**" to knock off Groupon or LivingSocial. 但是它没有对**优惠券**的关注**或关系**击败高朋团购或 LivingSocial 的**当地企业**. （and/or 导致的结构歧义）
			其它结构歧义	3	The death toll from Wednesday's storms seems **out of a bygone era**, 星期三的暴风雨的死亡人数**由于远古时代**好像发生了, （包括介宾结构做表语被误分析为状语 2 例，双宾结构 1 例）
			主语相关	2	It's smart for Facebook **to emphasize social experiences rather than just compete to offer the deepest discounts**, 对 Facebook 网站(脸谱网)来说是聪明伶俐的**强调社会经验而不是[刚刚;只是]竞争提供给高折扣的**, （包括主语从句、主语的补充说明 2 例）
			其它	4	Firefighters searched **one** splintered pile **after another** for survivors Thursday, 消防人员在(在)星期四**另外一个**幸存者**之后**搜索一裂成碎片的[绒毛;堆], （固定搭配 one…after another）

表九 系统融合系统在语料 D 上“词序不对”错误的分类统计

分类	频次	比例	具体分类	频次	举例及说明（第一行为原文，第二行为机器译文，第三行括号内容为错误说明）
定语	32	0.46	介词结构	16	These studies have gotten out of hand: This whole social-science trend **of analyzing pop hits** is ridiculous, 这些研究失控:整个社会科学的发展趋势**分析的流行歌曲**是荒谬的, （这类错误包括“of 结构”6 个）
			定语从句	8	But at least he's "challenging the cultural assumptions **that cause a lot of people to make bad life decisions**." 但至少他的“挑战文化假设,**引起了许多人的生活使坏的决定**。 ”
			不定式	2	Grad school has become a socially acceptable way **to drink beer, read, and go into massive debt in your 20s**. 研究生院已经成为一种社会所接受的方式**来喝啤酒,读书,走进 20 多岁的时候你的巨额债务**。
			过去分词	2	becoming the latest competitor to enter an already "crowded market" **dominated by**

					Groupon and LivingSocial. 成为最新的竞争对手进入一个已经“拥挤的市场”**由 Groupon 和 LivingSocial 为主**。 （dominated by …）
			其它	4	said Lim Boon Teck, Manager **Corporate Sustainability at HSBC Brunei**. 林布恩 Teck 说,**公司**经理**在汇丰文莱的可持续性**。 (参考译文：香港上海汇丰银行文莱分行的企业可持续性经理林文德说。)
状语	27	0.39	介词结构	17	One family rode out the disaster **in the basement of a funeral home**, 一个家庭安然渡过了灾害**在地下室的另一家殡仪馆**,
			状语从句	4	That's because people are often less inhibited and less self-conscious **when they're in dimly lit places** 这是因为人们往往不拘束,不自觉的,**当他们在灯光昏暗的地方**
			其它	6	Facebook Deals launches **tonight** & Groupon doesn't stand a chance “ Facebook 交易推出**今晚**& Groupon 不站在一个机会”
			时间	10	略
			地点	2	略
			其它	15	略
其它	10	0.14			She reported without incident for nearly an hour **before** her interpreter heard words in the Arabic-speaking crowd that gave him pause. 她没有事故报道了将近一个小时**之前**她的翻译说阿拉伯语的人听说他停顿了。 (此类错误含 10 个例子，错误性质各不相同，比较分散)

表十 规则系统与系统融合系统在语料 D 上“词序不对”错误的长度及距离统计

	规则系统			系统融合系统		
	短距离	长距离	总	短距离	长距离	总
单词	7	3	10	8	11	19
词串	8	48	56	20	30	50
总	15	51	66	28	41	69

从表八、九和十可以计算和推断出：

1）“词序不对”问题主要来源于翻译的源语言和目标语言的语序不一致。英语的介词结构、从句、不定式等作修饰成分时一般位于中心词后面，而汉语的各种修饰成分一般位于中心词前面，这是英汉翻译“词序不对”问题的主要原因。本文中，规则式系统和系统融合系统的“状语和定语造成的词序问题”分别占各自词序问题的 65%和 85%；

2）“状语和定语造成的词序问题”有相当一部分来源于介词结构和从句。在规则系统中，有 27%的词序问题来源于介词结构，17%来源于从句；在系统融合系统中，有 48%的词序问题来源于介词结构，17%来源于从句；

3）与统计系统相比，规则系统“状语和定语”之外的其它原因造成的词序问题也比较突出，占全部词序问题的 35%，通过分析可以看出，这些问题规律性较强，主要集中在各种原因导致的结构歧义、命名实体词组、固定搭配的翻译等问题上；

4）“词序不对”错误的长度及距离的统计表明：1’每个系统中，单个词的词序错误的总和远小于词串的词序错误的总和；2’词串的长距离词序错误比较常见，这一点在规则系统中表现得尤为明显；3’规则系统中单个词的短距离错误多于长距离错误，融合系统与之相反，单个词和词串的长距离错误均多于短距离错误；

5）在导致语序问题的状语中,“时间、地点和其它”的比例在两个系统中很接近,分别为 10:2:13 和 10:2:15。导致这一现象的原因，是否与语料本身有关，还有待进一步的研究确定。

4.4 译文与原文意思相反

由于“译文与原文意思相反”这类错误出现的概率比较小，所以作者以英汉项目整个测试集（500 句）作为样本，重点考察了规则式系统和统计式系统（层次短语系统 A）被判断有此类错误的语料，分别是 41 句（981 个单词）和 62 句（1256 个单词）。考察的结果请参见表十一和表十二。

表十一 规则式系统被判断为“译文与原文意思相反”的译文情况统计

判断情况	句错率	具体类型	句数	句错率	举例及说明（第一行为原文，第二行为机器译文，第三行括号内容为错误说明）
正确判断	0.00	译文与原文意思相反	1	0.00	While worksheets are considered soul sapping in the schools where many Junior Kumon parents **aspire** to send their children -- the kind of affluent places holding screenings of "Race to Nowhere" -- at Kumon, they are the essence of the experience. 当工作表被认为是在接受学位考试逐渐**削弱**许多年轻的 Kumon 的父母**渴望**派他们的孩子[(在)哪里;那里]的心灵的时候--这种富裕的地方[拘押;拥有;拿着;进行]跑到的[筛选;碎渣]不存在的地方--在 Kumon 那里,他们是经历的本质.
误判		张冠李戴	4	0.01	But asked whether Kadhafi was still alive, the Italian NATO general said: "We don't have any evidence. We don't know what **Kadhafi** is doing right now." 但是问卡扎菲是否还健在,意大利北约将军说:我们没有任何证据.我们现在不知道**自己**在干什么..
		主动被动颠倒	1	0.00	He denies **her** kidnap and murder. 他否认**她的**绑架和谋杀. 参考答案：他否认绑架并谋杀了她。
		主宾颠倒	1	0.00	Because despite the inexorable toll children **take** on our finances, our patience, our emotions and our energy reserves, they really are pretty great. 因为尽管孩子**承担**我们的财政,我们的耐心,我们的感情和我们的能源储备的无情的[通行费;钟声;伤亡],他们确实是非常棒的.
		否定移位	1	0.00	Having children introduces you to things you **never previously** cared about: soccer and swim team and peace signs as a fashion statement, pandas and airline insignias and rock-collecting. 有孩子把你介绍给你**从不以前**关心的事情:足球和游泳队和 V 形手势[随着;作为;例如]一个时尚宣言,熊猫和航空公司徽章和收集石头.
		其它误判	33		I don't know if scientists have looked into whether parents smile and laugh more than **non-parents**, but I'll bet they do. 我不知道科学家是否研究父母是否笑超过**不重要的父母**,但是我将断定他们做. (错译等)

表十二 层次短语系统被判断为“译文与原文意思相反”的译文情况统计

判断情况	句错率	具体类型	句数	句错率	举例及说明（第一行为原文，第二行为机器译文，第三行括号内容为错误说明）
正确判断	0.03	译文与原文意思相反	15	0.03	Five steps to going **vegan** 五步骤要吃**荤** (漏译、词语对齐错误造成)
		张冠李戴	6	0.012	If "**America** Can Do Whatever **We** Set Our Mind To," How Come Our Leaders Won't Set Their Minds on Jobs? 如果**我们**下定决心把“**美国**做什么就做什么，为什么我们的领导人不会安心工作吗？
		主动被动颠倒	3	0.006	Are you going to eat it? a girl **interrupted**. “你要吃呢？ ”一个女孩**被中断**。
		时序颠倒	3	0.006	**A day after the wedding,** the newlyweds asked the media not to intrude on their first weekend of married life, which they spent at home **before** William returned to military duty. **此前一天，在婚礼上，**新婚夫妇要求媒体不要打扰他们婚后的第一个周末，他们在家里度过**前**威廉回到军事义务。
		因果颠倒	2	0.004	The "more intelligent children in both studies grew up to drink alcohol more frequently and in greater quantities than less intelligent children," says Liz Day at Discover. 这两项研究“更聪明的孩子在长大，喝酒更频繁和更多的孩子**更聪明**， ”皮克说天发现。
		主宾颠倒	1	0.002	**Talk to vegans** about the ins and outs, get onto vegan blogs, acquaint yourself with vegan products, and study the health implications. **素食者谈谈**的来龙去脉，了解自己与严格的素食主义者博客产品，和健康的影响的研究。
		数字错乱	1	0.002	Based on HSBC research, global Climate Change related business or Green Business could be worth **US$2 trillion** ($2.45 trillion) by 2020. ”汇丰银行研究的基础上，全球气候变化相关的业务或绿色商业价值 **1.2 千亿美元**（ 2.45 万亿美元）到 2020 年。
		年岁错乱	1	0.002	She's served as queen for 59 **years**, having ascended the throne when she was just 25. 她担任女王为 59 **岁**，登上了王位时，还只是 25 。
		否定移位	1	0.002	Having children introduces you to things you **never previously** cared about: soccer and swim team and peace signs as a fashion statement, pandas and airline insignias and rock-collecting. 有孩子的人，把你介绍给你所**不**知道的东西**以前**关心：足球和游泳队和和平的迹象，一个时尚、熊猫和航空公司肩章和收集。
		其它错判	29	0.36	Businesses can't keep **cutting** their prices to attract new customers, so many daily deal services will soon "start to crumble." 企业不能把他们的价格来吸引新顾客，太多的日常服务将很快开始崩塌。 ” (包括实词缺失、错译等)

从表十一和表十二可以看出：

1）“译文与原文意思相反”的误判率很高，其中规则式系统的误判率为（41-1）/41=97.6%，层次短语系统的误判率为（62-15）/62=75.8%；

2）在“译文与原文意思相反”这个类型上，统计式系统的句错率（3%）明显高于规则式系统(0%)，造成这一现象的原因很可能是统计式系统在翻译过程中出现了词语对齐错误、漏译等问题；

3）在误判的各种类型中，也有比较集中的几种错误子类，包括：张冠李戴、主动被动颠倒、时序颠倒、因果颠倒、主宾颠倒、数字错乱、年岁错乱、否定移位等等，其中“张冠李戴”的错误比较明显，在两个系统中均达到了1%的句错率。

4.5 以上几种错误类型的基本结论

以上几种错误类型的基本结论可以归纳为表十三的内容。

表十三 英汉几种错误类型分析的基本结论

错误类型	译文选词错误	实词译文缺失	词序不对	译文与原文意思相反
出现概率	5%~8%(词错率）	1.4%~1.8%(词错率）	47%~51.8(句错率）	0%~3%(句错率）
产生原因	词的歧义(主要是名词和动词的歧义）占大多数，其次是固定搭配。	命名实体缺失的问题比较突出	英语中后置的定语和状语（较多来源于介词结构和从句）	
系统差异	规则系统比统计系统“消歧错误”和“兼类错误”更多，而“固定搭配错误”较少。	统计系统“译文消失”的词错率较高	规则系统的其它词序问题规律性强；统计系统词串的长距离词序错误比规则系统少。	统计式系统的句错率明显高于规则式系统
其它结论			词串的长距离词序错误较常见	

5 结论与展望

本文在 CWMT2013 人工评测的基础上，对几种主要的错误类型展开了进一步的抽样分析，得出如下结论：

1）译文选词错误、实词译文缺失和词序不对是目前机器翻译存在的主要问题；

2）造成以上问题的主要原因包括：词（主要是名词和动词）的歧义的存在、命名实体缺失、目标语言与源语言的语序不一致（如英语的后置定语和状语与汉语语序不一致）、固定搭配的处理不当等；

3）通过对比规则式系统和统计式系统，作者发现：规则式系统的优势是“固定搭配”导致的译文选词错误较少，“译文与原文意思相反”的错误极少，没有发现“译文消失”错误；统计式系统的优势是“消歧错误”和“兼类错误”导致的译文选词错误较少，词串的长距离词序错误相对较少。

由于本文主要是针对英汉翻译项目进行的错误分析，虽然机器翻译存在的问题大都是相通的，但翻译方向不同，各个错误类型上的表现也会不尽相同，比如，英文词的歧义现象比汉语更严重，所以英汉和汉英机器翻译系统在“译文选词错误”上很可能会表现出数量的不同。另外从本文也可以看出，采用不同的翻译方法或模型，在错误类型上的表现也会有所不同。即便是采用同一种翻译方法或模型，翻译系统本身的技术特点等因素也会影响其在不同的错误类型上的具体表现。以上种种原因再加上抽样语料本身的内容、样本规模（包括所选的系统和语料数量)的限制以及评测者个人的主观因素，使得本文中出现的词错率和句错率等数据仅具有相对的参考意义，得出的结论也不一定完全符合客观事实，故恳请细心的读者们甄别使用。

本文主要分析了机器翻译中几种常见的错误类型以及产生这些错误的语言学原因，作者认为这些其实也是机器翻译的难点问题。本文对机器翻译系统目前为什么没能更好地避免这些错误、以及是否有更好的方法来回避这些错误没有进行更深层的探讨，这也是我们后继研究工作需要重点关注的地方。

致谢

感谢 CWMT2013 评测尤其是人工评测参评单位同行们的支持。感谢三位匿名审稿人和孟凡东同学对本文所提的意见。感谢新疆大学麦热哈巴老师对本文涉及的维语问题的解答。感谢吕雅娟老师、姜文斌老师、谢军博士、于惠博士、杨似彤等同学对 CWMT2013 评测工作的贡献。

参考文献

Ariadna Font Llitj´os, Jaime G. Carbonell, and Alon Lavie. 2005. A framework for interactive and automatic refinement of transfer-based machine translation. In Proc. Of the 10th Annual Conf. of the European Association for Machine Translation (EAMT), Budapest, Hungary,May.

David Vilar, Jia Xu, Luis Fernando D'Haro, Hermann Ney. 2006. Error Analysis of Statistical Machine Translation Output. Fifth Int. Conf. on Language Resources and Evaluation (LREC), pages 39–45, Athens.

赵红梅，谢军，吕雅娟，于惠，张昊亮，刘群. 2013. 第九届全国机器翻译研讨会（CWMT 2013）评测报告. http://nlp.ict.ac.cn/Admin/kindeditor/attached/file/20140310/20140310173732_36859.pdf .