TIMSS 试题的机器翻译系统构建及其效果

蓝杨

（浙江警官职业学院，浙江杭州，310018）

摘要：研究以范例为基础的机器翻译技术和英汉双语对应的结构辅助英汉单句语料的翻译。翻译范例是运用一种特殊的结构，此结构包含来源句的剖析树、目标句的字符串、以及目标句和来源句词汇对应关系。将翻译范例建立数据库，以提供来源句作词序交换的依据，最后透过字典翻译，以及利用统计式中英词汇对列和语言模型来选词，产生建议的翻译。研究是以 2010 年国际数学与科学教育成就趋势调查测验试题为主要翻译的对象，以期提升翻译的一致性和效率。以 NIST 和 BLEU 的评比方式，来评估和比较在线翻译系统和本系统所达成的翻译质量。

关键词：试题；语料；剖析树；自然语言；机器翻译；TIMSS

中图分类号：H315.9 **文献标识码**：A **文章编号**：1672-3104(2013)05−0244−08

一、引言

国际教育学习成就调查委员会(The International Association for the Evaluation of Education Achievement, 以下简称 IEA)的主要工作是了解各国学生数学及科学(含物理、化学、生物、及地球科学)方面学习成就、教育环境等影响学生学习成效的因素，找出关联性，并在国际间相互作比较。自 1970 年起开始第一次国际数学与科学教育成就调查后，世界各国逐渐对国际数学与科学教育成就研究感到兴趣，IEA 便在 1995 年开始每四年办理国际数学与科学教育成就研究一次，称为国际数学与科学教育成就趋势调查(Trends in International Mathematics and Science Study，以下简称 TIMSS)。

中国教育科学研究院于 1983 年正式成为 IEA 的团体会员，并计划加入和引进 TIMSS 的调研活动，以期对中国数学教育和科学教育产生积极的作用。而我国的台湾省于 1999 年加入 TIMSS 后，已经开始着手实施相关工作，包括负责试题翻译及测验工作。本文在对国外和台湾的相关试题测试工作进行研究和分析后，对 TIMSS 试题翻译作了初步的研究分析。

以往使用人工翻译虽然可以达到很高的翻译质量，但是需要耗费相当多的人力资源和时间，而且在翻译过程中不同的翻译者会有不同的翻译标准，相同的翻译者也可能在文章前后翻译方式不一致而产生语意上的混淆。因此此类语言转换导致的问题间接影响试题难易程度。若直接将英文词汇透过英汉字典翻译成相对的中文词汇，翻译的结果可能会不符合一般人的用词顺序。另外中文的自由度较高，很容易造成翻译上用词顺序的不同。例如：“下图显示某一个国家所种谷物的分布图”，也可翻译为“某一个国家所种谷物的分布图，如下图显示”。可能会影响到受测者的思绪，使作答时粗心的情形增加。因此，若能利用机器翻译(machine translation)的技术来辅助翻译以及调整词序，便可提高翻译的质量和效率。

Dorr 等学者[1]将现在机器翻译依据系统处理的方式来分类，分成以语言学为基础翻译(linguistic-based paradigms)，例如基于知识(knowledge-based)和基于规则(rule-based)等；以及非语言学为基础翻译(non-linguistic-based paradigms) ，例如基于统计(statistical-based)和基于范例(example-based)等。

以知识为基础的机器翻译(knowledge-based machine translation)系统是运用字典、语法规则或是语言学家的知识来帮助翻译。这种利用字典来帮助翻译的系统，会有一字多义的情形发生，一个词汇在字典中通常有一个以上的翻译。以英翻中为例“current”这个字在字典里就有十多种不同的翻译，即使专家也无法找出一个统一的规则，在何种情况下要用何种翻译，所以在翻译的质量和正确性上很难满足使用者的

收稿日期：2013−01−07；**修回日期**：2013−09−15

作者简介：蓝杨(1980−)，男，畲族，浙江临安人，浙江警官职业学院副教授，主要研究方向：英语语言文学.

的需求。因此，翻译系统通常都会限定领域来减少一字多义，例如“current”在电子电机类的文章中出现，最常被翻译为电流，在文学类的文章中，最常被翻译为现代。

以范例为基础的机器翻译(example-based machine translation，以下简称为 EBMT)的相关研究已有相当多年历史，在 1990 年美国学者 Brown 和 Pietra[2]所提出的 EBMT 是将翻译过程分为分解(decomposition)、转换(transfer)和合成(composition)三步骤。分解阶段是将来源句放到范例库中搜寻，将所搜寻到 word-dependency tree 当作来源句的 word-dependency tree，并且形成来源句的表示式；转换阶段将来源句的表示式转换成目标句的表示式；合成阶段将目标句的表示式展开为目标句的 word-dependency tree，并输出翻译结果。Al- Adhaileh 等学者[3]将 structured string tree correspondence(SSTC)运用在英文翻译成马来西亚文的过程中，SSTC 是一种能将英文对应马来西亚文的结构，但此结构并没有解决词序交换的问题。目前较完整的 EBMT 系统为 tree-string correspondence (TSC)结构和统计式模型所组成的 EBMT 系统[4]，在比对 TSC 结构的机制是计算来源句剖析树和 TSC 比对的分数，产生翻译的是由来源词汇翻译成目标词汇的机率和目标句的语言模型所组成。

我们提出双语树对应字符串的结构(bilingual structured string tree correspondence，简称为 BSSTC)是可以运用在多元剖析树上的，并且 BSSTC 可在翻译过程中当作词序交换的参考。根据我们实验结果，我们能有效的调动词序，以提升翻译的质量。完成词序交换后，再透过字典翻译成中文，最后运用统计式选词模型，产生初步翻译结果，但本系统尚属于半自动翻译系统，故需要人工加以修饰编辑。

二、系统架构

由于我们的目的在于利用中英互为翻译的句子找出词序关系，并且将英文句和中文句词序的信息储存在计算机中，储存的格式是将中英文句的词序关系记录在英文剖析树的结构中，此结构将成为之后英文句的结构调整为适合中文的结构的参考。最后再将英文词汇翻译成中文词汇，并利用统计式选词选出最有可能翻译成的中文词汇，让翻译的结果更符合一般人的用词和顺序。

本系统的架构如图 1 所示。我们针对范例树产生的系统和英文句翻译系统这两部份分别简介如下。

范例树产生系统：这个系统利用中英平行语料作为基础，这里的中英平行语料必需要一句英文句对应一句中文句，且每一组中英文句都要是互为翻译的句子。中文句经过断词处理后，被断成数个中文词汇，以空白隔开；英文句则经过英文剖析器建成英文剖析树。将断词后的结果和英文剖析树经过剖析树对应字符串模块处理，建成英文剖析树对应字符串的结构树，此结构树称为范例树。再将每个范例树取出子树，并且判断是否有词序交换，将需要词序交换的范例树全部存入范例树数据库中方便搜寻。

英文句翻译系统：当输入英文句后，先将句子透过英文剖析器，建成英文剖析树。有了英文剖析树就可以透过搜寻范例树模块，标记英文剖析树上需要调动词序的结构，并依照所标记的词序作调整。词序调整完成后再将英文结构树中的英文单字或词组透过翻译模块做翻译。其中翻译模块包含了大小写转换、断词处理和禁用词过滤等环节，之后将处理过的词汇透过字典文件做翻译[5]。每个英文单字或词组都可能有

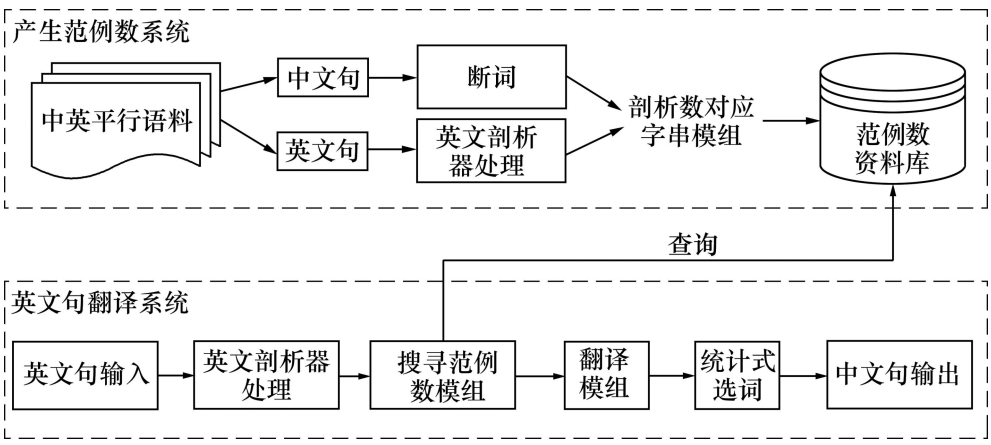

图 1　系统框架图

一个以上的中文翻译，因此需要选词的机制来产生初步翻译结果，此翻译结果尚需要人工作后续的编修。

三、系统相关技术

根据上一节介绍，系统架构分为范例树产生系统和英文句翻译系统两大系统。范例树产生系统的执行流程为先处理中文句断词和剖析英文句，再将断词和剖析后的结果输入至剖析树对应字符串模块，并将处理后的范例树存入数据库中。英文句翻译系统的执行流程区分为三大部分，第一部分为搜寻范例树模块，将英文剖析树跟范例树数据库作比对，并且将未比对到的子树做修剪；第二部分将修剪后的剖析树输入到翻译模块翻成中文；第三部分以中英词汇对列工具及 bi-gram 语言模型，计算出中英词汇间最有可能之翻译组合。

(一) 双语树对应字符串的结构(BSSTC)

在建立 BSSTC 结构之前，我们必须将中英平行语料中的中英文句先作前处理，我们将英文句透过 StanfordLexParser-1.6[6]建成剖析树，剖析树的每个叶子节点为一个英文单字，并以英文单字为单位由 1 开始标号。这里我们将树根定义为第 0 层，树根的子树是第 1 层，越往下层数越大，故叶子节点必定是英文单字，且不属于任何一层，如图 2 所示。中文句子断词后的单位由 1 开始标号。这里的中文句代表来源句；英文句则代表目标句。本结构都假设中英文对应是在词汇的对应或连续字符串的对应基础上。假设剖析树的节点集合 $N=\{N_1, N_2, \cdots, N_m\}$，$m$ 为剖析树上节点个数，对任一节点 $n \in N$，n 有三个参数分别是 n[STREE//]、n[/STC/]和 n[//ORDER]；我们以 n[STREE/STC/ORDER]来表示。为了方便说明，若节点 n 只有 n[STREE//]和 n[/STC/]，则以 n[STREE/STC/]表示。再假设 $nC(n)$为节点 n 有 1 到 $C(n)$个子节点。n[STREE//]为节点 n 所涵盖来源句的范围，层数最大节点的 n[STREE//]必定对应到一个来源句单字，此参数的功用为当作每个节点的键值(primary key)，故在同一棵剖析树中 n[STREE//]不会重复。图 3 是一个 BSSTC 结构的例子，来源句为英文：“Our experiments were simple in concept”；目标句为中文：“我们的实验概念很简单”。首先英文句必须先建成剖析树，每个叶子节点为一个英文单字，并以英文单字为单位做标号，例如：“Our(1)”，“ex-periments(2)”，“were(3)”，“simple(4)”，“in(5)”，“concept(6)”。另外中文句经过断词的处理后，以断词后的单位做标

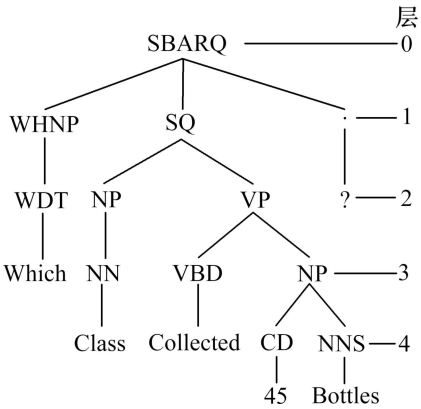

图 2 英文剖析数

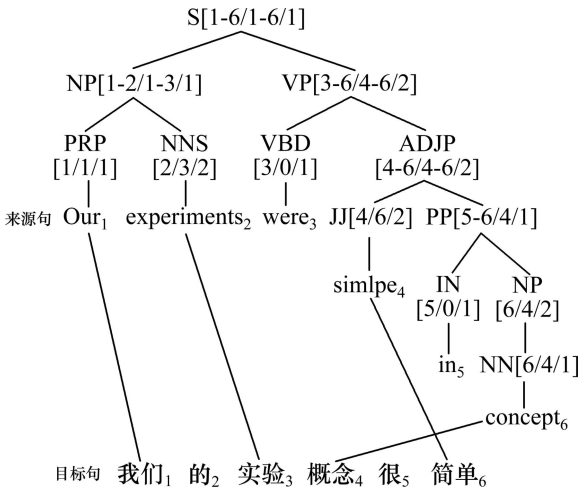

图 3 BSSTC 结构的表示法

号，例如：“我们(1)”，“的(2)”，“实验(3)”，“概念(4)”，“很(5)”，“简单(6)”。中英对应句都标号后，以标号为单位开始做词汇对准(word alignment)，并标记在剖析树的节点上。剖析树是用文法结构来分层，不同层节点能对应到不同的范围的目标句字符串。n[STREE/STC/]若为 VP[3-6/4-6/]，则 STREE 代表节点 VP 对应来源句第三到第六个字 “were simple in concept”；STC 代表“were simple in concept”对应目标句的第四到第六个字“概念很简单”。$nC(n)$[STREE/STC/ORDER]的兄弟节点(sibling node)若为 JJ[4/6/2]和 PP[5-6/4/1]，我们可以观察到 JJ 的 ORDER 大于 PP 的 ORDER，故 PP[5-6/4/1]的中文对应「概念」在 JJ[4/6/2] 的中文对应「简单」之前。

(二) 建立 BSSTC 结构和产生范例树

建立 BSSTC 结构必需要有英文跟中文互为翻译的句子，建构的顺序是从最底层也就是层数最大的开始标记，再一层一层往上建置到第 0 层为止，标记参数顺序是先将所有节点的 n[STREE//]和 n[/STC/]标记

完后，再标记 *n*[//ORDER]。首先，标记最底层 *n*[STREE//]的方法，是将最底层的节点 *n* 所对应叶子节点的编号标记在 *n*[STREE//]。如图 3 节点 NNS 所对应来源句的“experiments”的编号为 2，故 NNS [STREE//]中的 STREE 标记为 2。接着标记最底层 *n*[/STC/]的方法是寻找中英对应句中互为翻译的中文词汇和英文词汇，也就是词汇对准。词汇对准若采用人工方式，则相当耗时费力，其本身也是一项困难的研究。如图 3 来源句的“experiments”在字典中的翻译有“实验”、“经验”和“试验”，将这三个中文翻译到目标句去比对，此例子将会比对到目标句第三个词汇“实验”，接着将目标句“实验”的编号标记在 NNS[2/STC/]中的 STC 上。最后将比对到的个数除以英文句单字的个数，称为对应率。最佳情况下是每个英文单字都有相对应的中文翻译，对应率为 1；最差的情况下每个英文单字都没有相对应的中文翻译，对应率为 0，所以对应率会落在 0 到 1 之间，值越大代表对应率越高。我们需要够大的对应率，才能认定为范例树。因此，需要定一个门坎值来筛选，根据实验结果当门坎值越高留下来的范例树越少，而门坎值越低会使翻译的质量下降。

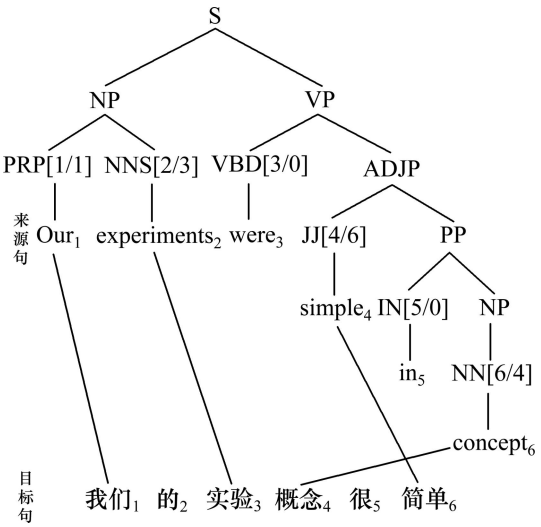

图 4　仅标记最底层

(三) 搜寻相同范例树

根据搜寻范例树算法的流程，如图 7。首先将来源句的剖析树加到数列(queue)里，从数列里面取出一棵剖析树到范例树数据库中，搜寻是否有相同结构的范例树；如为否，则将此棵树的下一层的子树加入数列，加入数列的顺序为左子树到右子树；如为是，则将该树的 ORDER 标记在来源句的剖析树上，继续取出数列内的剖析树，直到数列里没有剖析树为止。所

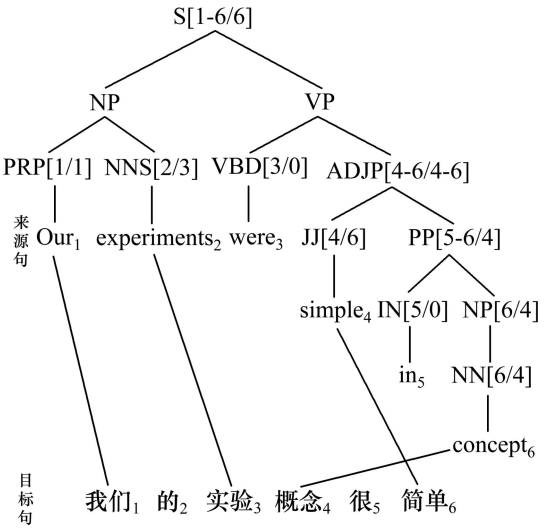

图 5　仅标记 STREE 及 STC

以来源句的剖析树是由一个以上的匹配子树所组成。

图 6 为剖析树搜寻范例树的情形。来源句：“The graph shows the heights of four girls”,剖析树为“(S(NP(DT The)(NN graph))(VP(VBZ shows)(NP(NP(DT the)(NNS heights))(PP(IN of)(NP(CD four)(NNS girls)))))”。透过搜寻范例树算法找出匹配子树，首先以节点 S 为树根的剖析树到数据库作搜寻，搜寻时不包含叶子节点，此例子没搜寻到匹配子树，则将节点 S 的子树 NP 和 VP 加入数列中。接下来将从数列中取出的子树为 NP，到范例树数据库搜寻匹配子树，但数据库中没有相同的范例树，此时 NP 的子树皆为叶子节点，所以并无子树在加入数列中。依照先进先出的原则下一个从数列取出的是 S 的右子树 VP，在范例树数据库中还是搜寻不到，因此要将 VP 的子树 VBZ 和 NP 加入数列中，但 VBZ 为叶子节点，故只有 NP 加入数列中。接下来是子树 NP 从数列中被取出来，子树 NP 在数据库中搜寻到相同的范例树，如图六的范例树就是所搜寻到的匹配子树，因此将范例树的 ORDER 标记上去，标记后的剖析树将如图 8 所示。此时数列中已经为空，搜寻范例树的流程到此为止。

标记完 ORDER 之后，将没有标记的子树作修剪，也就是将不用作词序交换的子树修剪到最小层树。如图 8 节点 S 的右子树、NP[2]和 NP[1]的子树皆不需要作词序交换，因此修剪的结果为“(S(NP The graph)(VP(VBZ shows)(NP(NP[2] the heights)(PP[1](IN[2] of)(NP[1] four girls))))) ”，如图 9 所示。最后从层数最大的每个兄弟节点开始逐层往上依照优先权顺序调整剖析树的结构；调整后的结果将会输入到翻译模块

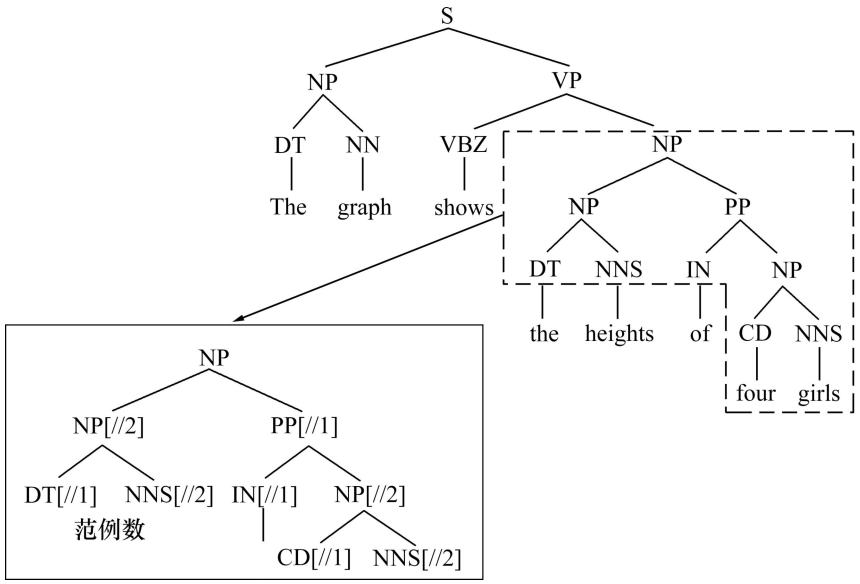

图 6 剖析数与范例数的对应关系

输入：来源句剖析数 S
范例数资料库 $D=\{D_1, D_2, \cdots, D_m\}$, $D_i \in D_i$ 包含 T_i 与 O_i, i 为 1 到 m
T_i 是第 i 棵范例数，O_i 是 T_i 所标记的词序
开始
设计列 Q 用来储存剖析数，初台为 NULL
S 加入 Q
当 $Q \neq NULL$
从 Q 中 pop 一棵范例数
如果在 D 中搜寻到相同的范例数 T_i
则将 Q 标注在 S 上
否则将下一层子数加入 Q
结束
输出：标记好 ORDER 的剖析数

图 7 搜寻范例数演算法

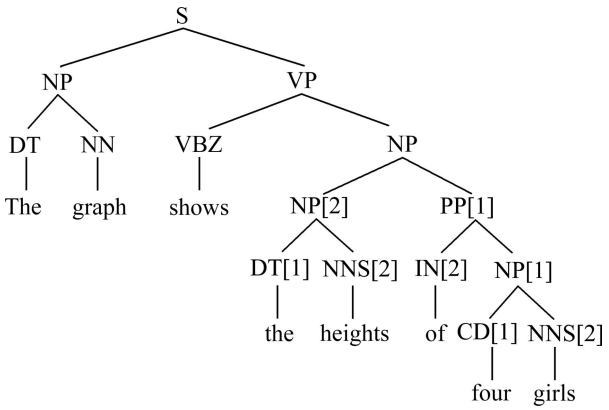

图 8 完成 ORDER 标记

产生翻译。若我们直接取来源句剖析树的叶子节点作翻译，将会成为单字式的翻译，我们将无法对词组或词组作翻译。翻译的部分会在下一节会作详细说明。

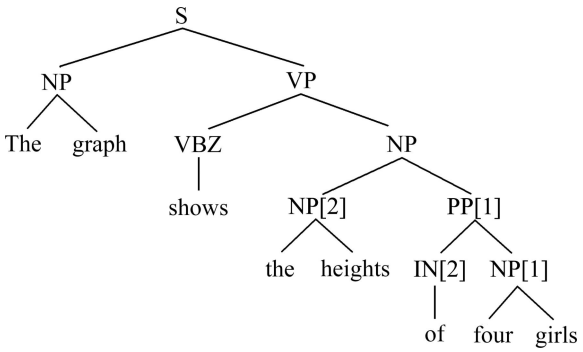

图 9 剖析书修剪后的结果

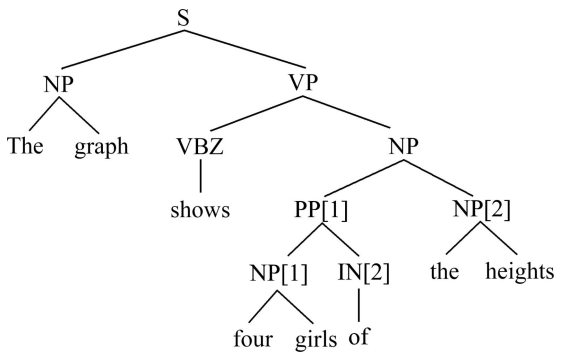

图 10 调整词序后的结果

(四) 翻译处理

经过上一节处理最后得到修剪树，修剪树的叶子节点可能为英文单字(word)、词组(term)。词组即为数个单字结合的字符串，不一定为完整的句子，如“would be left on the floor”或词组(phrase，如名词词组、动词词组、形容词词组等)，如“in order to”。在翻译处理上会遇到英文单字或词组，在英文单字的

部分，直接查寻字典文件作翻译；词组的部分利用规则词典文件的词组，和词组进行字符串比对，以找出符合的词组及中文翻译。以下为字典文件及规则词典文件分项说明。

字典文件：字典文件部分我们使用 Concise Oxford English Dictionary[8](牛津现代英汉双解词典，收录 39429 个词汇)，将前处理过后的英文单字或词组做翻译对等字搜寻的动作，找出所有和该英文单字的中文词组，作为翻译的候选名单。如无法在字典文件中搜寻到对应的中文翻译。如姓名和专有名词，则直接输出该英文字。

规则词典文件：为常用的名词词组、动词词组、形容词词组等词组，以及试题翻译小组所决议之统一翻译词组以人工的方式建立的中英翻译对照档，如 in order to(为了)。分成单字和词组翻译是因为若在规则词典文件比对不到，则用空白来做一般字和字之间的断词，也就变成单字的翻译，因为词组较能完整表现出动作或叙述。如只用单字作翻译，会造成翻译上的错误。须注意的是比对的句型若有相似结构但不同长度的字符串样式，则取长度最长的为结果。如一英文句子为“…as shown in diagram…”，同时满足规则词典文件内的“as shown in diagram”和“in diagram”片语句型，则我们会选择长度较长的“as shown in diagram”而不是选择“in diagram”加上“as show”作为断词的结果。在英文翻译成中文的过程中，有些英文单字不需要翻译或是无意义的情形，所以我们将这些单字过滤不翻译，这些单字称为 stop word。例如：冠词 the 直接去除。介词 for、to、of 等，若前一单字为 what、how、who、when、why 等疑问词，则允许删除，另外，to 出现在句首直接删除。助动词 do、does 等，判断方式与介词相同。在翻译过程中还可能出现词干变化(如~ing、~ed 等)和词性变化(如动词 break，其过去式为 broke，被动式为 broken，以及名词单复数型态)。词干变化的部份，我们可以还原各词性(名词、动词、形容词、副词)；词性变化的部分，有些是不规则的变化，较难用算法处理。

四、系统翻译效果评估

本节主要介绍利用本系统翻译国际数学与科学教育成就趋势调查 2010 年考题，简称 TIMSS2010，并将试题依照年龄别和科目别，分别比较翻译的质量。最后将与在线翻译以及已经研发在用的翻译系统作比较。评估方式为利用 BLEU(IBM 公司的机器翻译评测标准)及美国国家标准与技术研究院 NIST(National Institute of Standards and Technology)指标。

(一) 实验来源

用来翻译的来源为 TIMSS2010 试题， 所有实验语料句对数、中英词汇数、中英总词汇个数及平均句长，皆如表 1 所示。用来建立范例树的来源有中国教育科学院委托北京实验二小和北京第四中学语文学习教科书补充资料题库[7]及科学人杂志。补充数据题库以人工方式完成中英语句对列(sentence alignment)，再经过范例树的筛选门坎值为 0.6 的情况下有 565 句。用来训练选词机率模型的来源有自由时报中英对照读新闻及科学人杂志。自由时报中英对照读新闻从 2009 年 2 月 14 日至 2011 年 10 月 31 日，而自由时报中英对照读新闻本身就已经作好中英语句对列。科学人杂志是从 2006 年 3 月至 2009 年 12 月共 110 篇为语料来源。

(二) 实验设计

首先，将 TIMSS2010 试题问句以逗号、问号或惊叹号作为断句的单位，每个诱答选项做为断句的单位，若一道题目为一句试题问句及四项诱答选项所组成，则一道题目可断出五句。经过人工断句处理 TIMSS2010 试题，小学数学领域有 165 句；小学科学领域有 262 句；中学数学领域有 439 句；中学科学领域有 236 句，并整理为文字文件。建立范例树数据库所使用的语料为中学补充数据题库，训练机率模型所使用的语料自由时报中英对照读新闻加上科学人杂志，其中训练语言模型得到的 bi-gram 共有 134435 个。

主要评估的对象有 Google 在线翻译、Yahoo 在线翻译及本系统互相做比较，并且评估翻译系统在不同年级的试题内容上，翻译质量是否会按照越低年级其翻译质量越好的趋势。因此，我们将实验组别分为中学生段和小学生段；数学领域以 M 为代号，科学领域以 S 为代号，当作实验组别的名称。可以 TIMSS2010 分为中学段 2010 M 组、中学段 2010 S 组、小学段 2010 M 组及以小学段 2010 S 组四组；在加上 TIMSS 2010 数学及科学领域之中学段试题，和 TIMSS 2010 数学及科学领域之小学段试题，分别为中学段 2010MS 组及小学段 2010MS 组，总共六组，如表 2 所示。

(三) 实验结果

从表 3 可观察到，中学段 2010 M 组 NIST 分数以 Yahoo!最高分，但 BLEU 分数与本系统相近，可知 Yahoo 对中学段 2010 M 组所翻译的词汇跟参考翻译较相同，但 Yahoo 和本系统翻译后词序的正确性是差不多的。小学段 2010 M 组试题中有较多特殊符号，例如○和●等，Yahoo 及 Google 在线翻译系统会将这

表 1 实验语料来源统计

语料	语言	句对数	词汇数	总词汇个数(tokens)	平均句长
中学补充资料题库	中文	2 059 句	2 333	12 460	6.1
	英文		2 887	13 170	6.4
科学人	中文	4 247 句	9 279	70 411	16.6
	英文		10 504	68 434	16.1
自由时报中英对照读新闻	中文	4 248 句	19 188	145 336	34.2
	英文		25 782	133 123	31.3

表 2 TIMSS 试题实验组别表

中学 2003M 组	中学 2003S 组	小学 2003M 组	小学 2003S 组	中学 2003MS 组	小学 2003MS 组
TIMSS2003 中学数学领域试题	TIMS2003 中学科学领域试题	TIMSS2003 小学数学领域试题	TIMSS2003 小学科学领域试题	TIMSS2003 中学数学及科学领域试题	TIMS2003 小学数学及科学领域试题

表 3 本系统及以上翻译系统之 NIST 及 BLEU 值比较表

组别	中学 2003M 组		中学 2003S 组		小学 2003M 组	
指标	NIST	BLEU	NIST	BLEU	NIST	BLEU
本系统	4.700 2	0.144 0	4.408 9	0.125 4	3.981 9	0.130 4
Google	4.526 8	0.146 7	4.858 7	0.184 8	3.757 3	0.101 6
Yahoo!	4.879 3	0.145 5	4.613 6	0.139 6	4.045 7	0.141 9
组别	小学 2003S 组		中学 2003MS 组		小学 2003MS 组	
指标	NIST	BLEU	NIST	BLEU	NIST	BLEU
本系统	4.222 8	0.101 8	4.861 3	0.130 9	4.440 0	0.113 8
Google	4.444 5	0.152 7	4.934 3	0.161 1	4.472 0	0.134 4
Yahoo!	4.436 1	0.144 2	5.075 5	0.143 5	4.607 0	0.143 6

些特殊符号处理成乱码，但本系统可以将特殊符号保留下来，故小学段和中学段 2010 M 组与最高分系统的差距较小。先前我们假设翻译质量是否会按照越低年级其翻译质量越好的趋势，观察中学段 2010MS 组及小学段 MS 组，可发现与假设相反，各系统在中学段 2010 MS 组的表现都比小学段 2010 MS 组要好。可推测出本系统其中一种语料为中学补充数据题库较符合 TIMSS 中学段 2010 的试题。

我们将中学段 2010M 组和中学段 2010S 组作比较，小学段 2010 M 组和小学段 2010 S 组作比较，可以发现各系统除了 Google 之外，在 M 组上表现都比 S 组好，因为 M 组的试题内容包含较多的数字，对于翻译系统较容易处理，而 S 组则包含较多专有名词，对于翻译系统较为困难。

五、结论

本论文提出 BSSTC 结构，此结构能够记录来源句词汇的位置、目标句词汇的位置及来源句与目标句词汇对应的关系；并且将 BSSTC 结构运用在我们实作的翻译系统上。本系统是利用 BSSTC 结构建立范例树，将来源句经过搜寻范例树算法，来达到修正词序的目的。最后，在依据修正后的词序进行翻译，翻译时再利用中英词汇对列工具及 bi-gram 语言模型，选出最适合的中文翻译，产生建议的翻译，此翻译还需要人工修整。 TIMSS 的试题为数学及科学类，应该要用大量数学及科学类的语料，但实际上我们并无法找到够多的数学及科学类语料，尤其以中英对应的语料最少，所以我们选用新闻及补充数据题库来拟补语料的不足。不过训练量还是不够多，在选词上会有许多机率为 0 的情况，造成选词错误。未来将尽量找寻相关领域的语料，来建立范例树和训练语言模型，就能针对不同领域的内容进行翻译，使翻译的结果更为精确。 训练语料中的断词是使用国外的系统，而我们翻译使用的字典为牛津字典，两者所使用的字典并不相同，会使断词后的词汇可能无法在牛津字典中找到，造成选词错误。未来可将翻译后的词汇，找出同

义词来扩充词汇数，便能增加被找到的可能性。

英文的语言特性上并没有量词，而中文句中运用了很多的量词，如缺少量词也会使中文的流畅度下将。本系统的翻译结果也缺少中文的量词。未来若能将翻译结果填补上缺少的量词，便可达到更好的质量，这也是我们今后要做的工作。

参考文献：

[1] B. J. Dorr, P. W. Jordan and J. W. Benoit. “A Survey of Current Paradigms in Machine Translation” Advances in Computers [M]. London: Academic Press, 1999: 1−8.

[2] P. F. Brown, J. Cocke, S. A. D. Pietra, V. J. D. Pietra, F. Jelinek, J. D. Lafferty, R. L. Mercer and P. S. Roossin. A Statistical Approach to Machine Translation [J]. Computa-tional Linguistics, 1990, 12(6): 79−85.

[3] M. H. Al-Adhaileh, T. E. Kong and Y. Zaharin, A synchronization structure of SSTC and its applications in machine translation [C]// Proceedings of the International Conference on Computational Linguistics-2002 Post-Conference Workshop on Machine Translation in Asia. 2002: 1−8.

[4] Z. Liu, H. Wang and H. Wu. Example-based Machine Translation Based on TSC and Statistical Generation [C]// Proceedings of the Tenth Machine Translation Summit, 2005: 25−32.

[5] 桂诗春. 标准化考试一理论、原则与方法[M]. 广州: 广东高等教育出版社, 1986.

[6] R．L．桑代克 E．P．哈根. 心理与教育的测量和评价[M]. 北京: 人民教育出版社, 1985.

[7] 藏忠恒. 心理与教育测量[M]. 上海: 华东师范大学出敝社, 1987.

[8] Bachman L F. Fundamental Considerations in Language Testing [M]. 上海: 上海外语教育出版社, 1999.

The Theoretical Basis and Empirical Research on Translation of TIMSS Testing Contents

LAN Yang

(Zhejiang Police Vocational Academy, Hangzhou 310018, China)

Abstract: The paper takes English-Chinese single sentence corpus translation as researching materials which are based on paradigms machine translation and English-Chinese corresponding auxiliary structure. Translation paradigm consists of parse tree, character string target sentence and source sentence. To build translation paradigm’s database as the basis for exchanging words sequence of source sentences and then produce ideal translation outcome via dictionary and statistical English-Chinese vocabulary. The empirical research takes TIMSS 2010 testing contents as the translation targets with the methods of NIST and BLEU in order to evaluate and compare online translation system and the system we are researching.

Key Words: test questions; corpus; parse tree; natural language; machine translation; TIMSS

[编辑：汪晓]