英汉双向规则主导型混合机器翻译系统

胡小鹏，袁 琦，耿鑫辉
中国电子信息产业发展研究院 北京 100048
E-mail：huxp@ccidtrans.com

摘要：本文介绍了我院研发的语言学知识模型与统计方法相结合的英汉双向规则主导型混合机器翻译系统的结构设计，其中包括从平行和可比语料库提取术语和翻译模板，从三元组可比语料库提取本族英语多词表达（MWEs）。文中给出了该实用型混合机器翻译系统的综合性能评价，例举了系统的典型应用，最后给出下一步工作设想。

关键词：规则主导型混合机器翻译；数据驱动的统计方法；语言学知识模型；混合术语提取

An English-Chinese Bi-Directional Hybrid Machine Translation System Guided by RBMT

HU Xiao-peng, YUAN Qi, GENG Xin-hui
China Center for Information Industry Development, Beijing 100048, China
E-mail：huxp@ccidtrans.com

Abstract：*This paper first reviews several typical techniques most commonly used and the most promising ones in the R&D of HMT guided by RBMT. It then gives a more detailed description of the various data-driven statistical approaches adopted by a practical English-Chinese bi-directional HMT system guided by RBMT that integrates linguistic knowledge models and statistical approaches developed by CCID. Such approaches include extracting glossaries, terminologies and translation templates from parallel and comparable corpora and extracting MWEs in native English from three-tuple comparable corpora. This paper also presents a comprehensive performance evaluation of this practical HMT system, illustrates typical applications of the system, and finally provides a vision for the future work.*

Keywords：*hybrid machine translation guided by RBMT， data-driven statistical approaches， linguistic knowledge models， hybrid terminology extraction approach*

1 引言

统计机器翻译已经有 20 多年的历史，其中大约前十年的时间统计机器翻译和规则机器翻译被视为互为竞争的范式。然而在后十年，人们对两种方法相结合的趋势产生越来越大的兴趣。这是因为纯统计的和纯规则的范式都有很强的局限性，同时也存在互补性。比如，对于短语和短距离搭配，统计系统的翻译往往是出奇的好，但是在选择长距离搭配词汇时它们常常失败，原因是基于 N-gram 的语言模型忽略了词汇的长距离搭配。相比之下，对于规则系统，尽管词汇选择较差，但是如果分析器对句子做出正确的分析，其输出往往是出奇的好[Reinhard et al., 2014]。另一方面，它们在应用上也有互补性，统计翻译系统在论坛、常见问答（FQAs）和用户生成内容（UGC）等社交媒体翻译方面具有优势，而规则系统在技术文档、报告、在线帮助和用户界面等翻译方面具有优势。甚至已有企业通过一套系统性的标准指南为用户确定翻译引擎最佳解决方案[Lori et al., 2013]。正如 2014 年 4 月哥德堡召开的混合机器翻译研讨会上 Reinhard Rapp 等人指出的，“鉴于统计和规则系统之间的互补性，它们间的界

限已经收窄，目前机器翻译领域正在出现一个新的跨范式的观点，培养基于统计和基于规则两种主要范式之间的创新组合将对现代机器翻译技术带来重大突破”[Reinhard et al., 2014]。

概括起来，目前有两种混合化发展趋势：或者把形态、句法或语义知识加入到统计系统，或者把数据驱动的统计方法与现有的基于规则的系统相结合[Marta et al.,2013]。考虑到上世纪80 年代以来产业界不断发展和积累的机器翻译语言学资源，在机器翻译产业应用与研发领域，采用现代统计技术融入现有规则系统的混合化方法，即以规则翻译为主导融合统计技术的混合机器翻译研发占有重要地位。

本文在相关工作中介绍了规则主导型混合机器翻译研发方面最常用且最有发展前景的几项典型技术和工程。在正文部分较详细介绍了我院研发的语言学模型与统计方法相结合的实用型英汉双向规则主导型混合机器翻译系统的结构设计、系统评价与应用。文章最后给出了下一步工作设想。

2 相关工作

近年来，为了应对新兴和迅速发展的科技领域词汇短缺以及平行语料库固有的时间滞后和特定领域文本稀缺问题，基于可比语料库的数据提取技术已成为规则主导型混合机器翻译研发中的一项基础技术。在可比语料库资源开发方面，自 1995 年 R. Rapp 提出并验证了“跨语言文本的词共现模式之间存在相关性”的假设以来，建造和使用可比语料库提取单词和多词表达的研究已受到业界高度关注。然而，几乎所有的方法都有一个共同点，即它们需要预先假定初始词典（即双语种子词典）以便找到两种语言之间的映射关系。与此相反，2014 年 R.Rapp 最新发展了他的“词共现模式相关性假设”，提出不使用种子词典，而是在对齐的可比文档中通过分析多词单位的每个组成成分的语境行为（contextual behavior），并组合其分析结果确定多词表达的翻译[Reinhard et al., 2014]。评估实验表明，利用这种不使用种子词典的可比语料库多词表达提取方法提高了由复合名词构成多词表达翻译的准确度。同时，由于该方法不使用种子词典，减少了对语言知识的依赖性，因此具有向其他语言对扩展的优势。在可比语料库应用方面，在欧盟第 7 科技框架计划（2007-2013）支持下，先后实施了用于机器翻译资源不足领域的可比语料库分析和评价（ACCURAT）项目以及基于可比语料库的术语提取（TTC）项目。两个项目的目的同样是解决目前机器翻译研发中平行数据资源短缺问题，尤其是特定和新兴领域（如再生能源）平行数据短缺问题。ACCURAT 项目重点放在制订可比语料库文档可比性度量标准以及可比语料库的词汇与术语对齐和提取方法[ACCURAT, 2012]，而 TTC 项目的目标是利用上述方法和工具从可比语料库自动生成欧洲语言及中文等双语词典[Béatrice et al., 2012]。

利用统计评价方法自动发现和定位规则系统中的潜在语法错误是提高规则主导型机器翻译性能的一项重要技术。机器翻译自动评价方法已被证明对监测机器翻译开发进展，统计机器翻译优化参数以及比较不同系统性能都是有效的，但是目前自动评价打分不提供发现典型翻译错误并做调试优先级分类的信息。从这个角度看，Bogdan Babych 认为机器翻译自动评价研究面临的挑战是发展一种按照词汇、语法和语体三个层面进行差异化和细粒度错误分析的研究方法[Bogdan et al., 2009]。他提出了利用自动评价方法自动发现和归类机器翻译中多词表达（MWEs）的翻译错误的方案，对于提高规则翻译系统质量，缩短研发周期很有实用价值[Bogdan et al., 2009]。该方案是在句对齐的平行语料库中自动检测高频多词表达并对每一个多词表达生成词语索引，计算自动评估得分，该得分表明某一特定的多词表达在语料库中是否系统性地被误译。该方法既可以应用到源语多词表达也可用到目标语多词表达，分别指出机器翻译系统是否能成功处理源语多词表达，或者某些高频目标语多词表达是否可以成功生成。所得到的结果对系统地检查机器翻译系统的覆盖范围是很有用的。

采用语言学模型与统计方法相结合的混合术语提取机制实现规则翻译系统领域适应性

解决方案也是一种有前途的规则主导型混合机器翻译的设计方法。早期的机器翻译领域适应性扩展和翻译质量提高是依靠人工开发特定领域词典并调试局部规则实现的。近年来，有很多文献发表了基于统计后编辑（SPE）方法的规则翻译系统领域适应性研究以及利用它改进商用机器翻译产品性能的报道[Pierre et al., 2007; Michel et al., 2007; Loic et al., 2009; Antonio et al., 2009; Terumasa et al.,2010]。相对于这种需要两种不同系统通过串行耦合方式实现的混合机器翻译结构，最近 Petra Wolf 等人提出使用传统语言学与数据驱动的统计方法交织结合的规则系统领域适应性解决方案[Petra et al., 2011, 2013]。如图 1 所示，实现该方法的数据流包括 4 个阶段，即术语获取、语言规则过滤、资源准备和新系统构成。在术语获取阶段，采用了独特的把规则翻译系统融入统计术语提取的混合术语提取机制。在语言规则过滤阶段采用多层语言规则过滤技术消除数据噪音。在资源准备和新系统形成阶段采用语言学属性值自动扩增技术将筛选并赋值的术语融入规则系统。从而，该方案在不需要人工介入的前提下自动实现规则系统的特定领域适应性能力。与目前采用单独的统计方法从语料库提取术语不同，混合术语提取机制的特点是仅提取翻译系统词典中没有出现的新词对作为候选。因此，该方法数据的 F-度量值高，把经过多层过滤的术语对融合到翻译系统后，避免了系统性能变差。

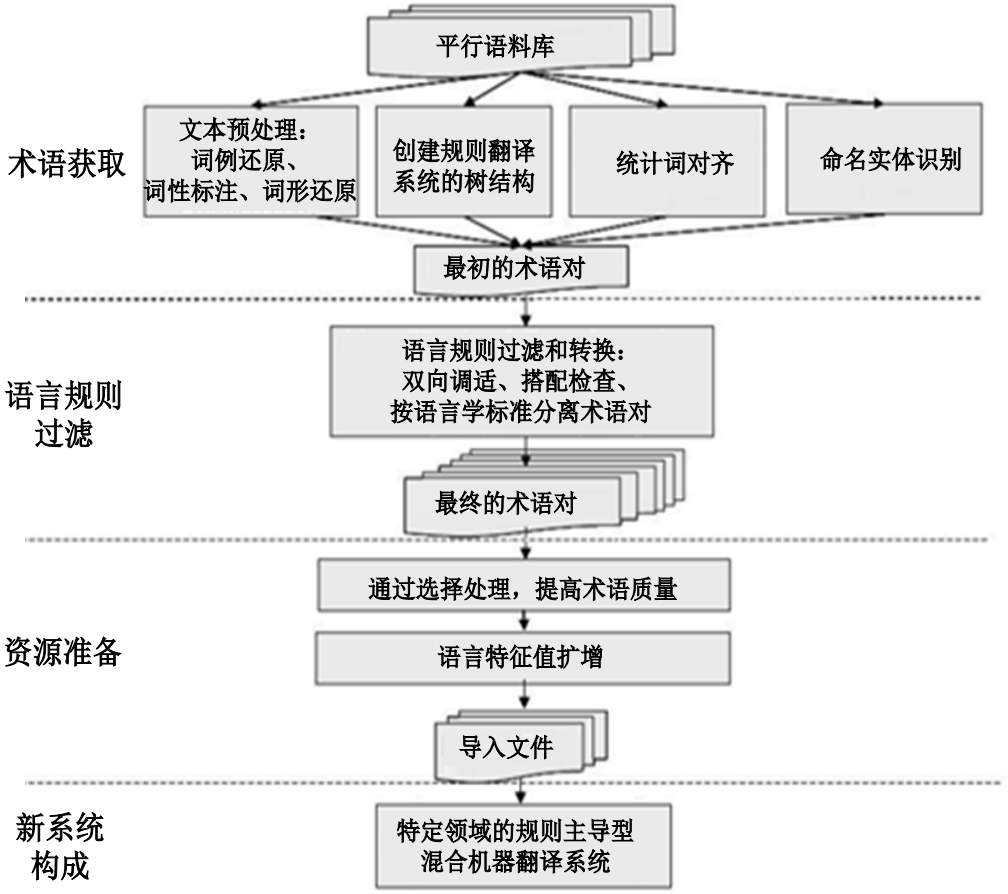


图 1 通过混合术语提取实现领域适应性流程

目前正在实施的有代表性的规则主导型混合机器翻译系统研发计划是纳入欧盟第七框架计划执行期为 2010-2014 年的高质量混合机器翻译（HyghTra）项目。该项目试图使用先进的统计方法通过提取平行和可比语料库双语资源，开发词典和语法规则库提高面向产业应用的规则主导型混合机器翻译系统的性能。当前文献中介绍的很多混合翻译系统已经尝试把某些基于语言学知识的分析性抽象放在统计机器翻译的核心之上的设计方法[Kurt et al., 2012]。Kurt Eberle 等人认为，这不是最好的选择，因为根据基本理念，统计机器翻译起初

是语言学无知的，它从语料库提取的信息通常是以庞大的数据集形式表现的，人类一般无法以自然的方式读取。这就意味着这种类型的系统难以提供接口以便用规范和简单的方式把语言学知识融入其中[Kurt et al., 2012]。因此，在把数据驱动的统计方法与语言学知识模型相结合的混合机器翻译研发中，HyghTra 项目采取了相反的做法 -- 把使用统计方法从语料库获取的信息最大限度地融合到规则翻译系统。如果项目中作为骨架的规则翻译系统在语言学上很成功并且结构上是充分模块化的，那么这样构成的混合机器翻译系统在实现高质量输出方面具有很大潜力。HyghTra 项目主要在以下 5 个方面融合了统计模型： ① 建造和使用可比语料库，从中提取平行资源；② 用 GIZA++与平行语料库扩展词典；③ 用语料库统计方法开发和维护机器翻译规则库；④ 利用自动评价技术评估系统最常用的语法结构和多词表达翻译质量；⑤ 通过从平行语料库挖掘正确的翻译，支持研发人员集中分析影响系统质量的关键语法结构[Kurt et al., 2012]。

3 英汉双向规则主导型混合机器翻译系统设计

组成该系统的核心系统是我院自主研发的实用型规则翻译系统，由于该核心系统有效集成了浅层句法分析器和翻译模板匹配等微引擎技术并采用规则与统计相结合的分词处理、命名实体识别和浅层句法分析技术，从而确保系统生成译文具有较强的句法结构和结构语义表达，这是研发本文提出的规则主导型混合机器翻译系统的重要前提。本文提出的混合翻译系统融合了以下 4 种统计数据资源，分别介绍如下。

3.1 从平行语料库提取术语

在本文提出的规则引导型混合机器翻译设计中采用统计与语言知识相结合的混合方法从平行语料库提取术语。通常双语术语提取的标准方法包括两个步骤，首先在源语言中识别术语候选，然后把源语言中的候选映射到目标语言的候选。本方法与标准方法不同之处在于直接通过构建短语表建立映射关系，然后通过多级过滤生成术语候选。由于减少了转换环节，提高了术语提取的正确率。如图 2 所示，本方法分为三个步骤，第一步，从平行语料库创建短语表；第二步，过滤术语候选从中选择符合语法规范的术语；第三步，对生成的术语列表由人工后编辑，确认术语的词形和属性标注的正确性。

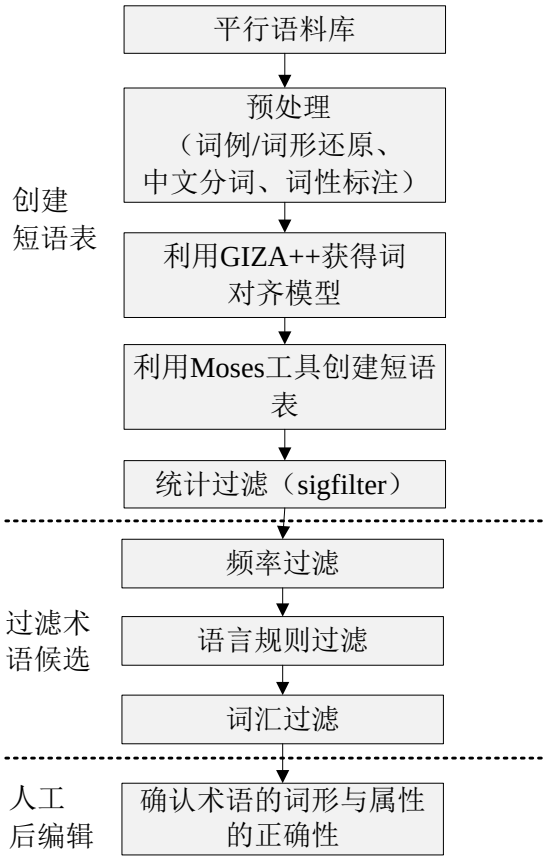


图 2 从平行语料库提取术语流程

3.1.1 创建短语表

在定制化开发中，使用用户提交的百万句对级平行语料库提取短语表。其流程包括：

① 预处理，使用工具对英语句子进行词例/词形还原、词性标注等，同样对中文句子进行分词/词性标注；过滤句对，去除原文/译文句子长度和原文/译文长度比值超过阈值的句对。

② 词对齐，利用 GIZA++处理平行语料，获得词对齐模型。

③ 短语表提取，利用 Moses 工具在词对齐模型基础上创建短语表。

④ 对短语表进行统计过滤（sigfilter）。

3.1.2 过滤候选术语

对候选短语进行以下三级过滤；

① 频率过滤器：只有频率和翻译概率高于给定阈值的短语被视为术语候选。

② 语言规则过滤器：设有一个内在的语言结构，只有符合这种结构的候选是合法的术语候选。根据术语结构的语言学规则，使用有限状态转移网络发现候选术语，例如在英语中，主要考虑三种模式的术语：Adj + N, N + N, N + Prep + N。由这三种模式扩展而形成的变体，也可以作为候选术语的筛选范围，如，multiple association measure（Adj + N + N）可以看成是由模式 Adj + N 和模式 N + N 扩展而成的[冯志伟，2008]。

③ 词汇过滤器：删除系统中已有的术语候选或是非特定领域的候选，以及包含停用词的术语候选。

对利用上述方法生成的英汉和汉英 4 个候选数据集（包括信息技术和体育领域），分别随机选择术语候选 1,000 条，通过人工评估，平均准确度分别达到 90.25%和 84.68%。

3.2 从可比语料库提取术语

从可比语料库提取术语流程如下：

① 建造可比语料库：根据我院研制的可比语料库文档可比性度量标准筛选强可比等级以上可比文档构建可比语料库。

② 分词：利用分词工具分别对中英文文本 C1 和 E1 进行分词处理并生成分词文件 C2 和 E2。

③ 生成词表：利用 GIZA++处理上述分词文件，分别生成中英文词表 CT1 和 ET1。

④ 词性标注：利用中英文分词及词性标注工具分别将可比语料库 C2 和 E2 文本进行词性标注。

⑤ 计算上下文向量：对于 C2 和 E2 中的每个词，选取其前后相邻的 ViewLen 个词作为其上下文向量的计算范围，相邻词中每个词与当前词的共现值乘以相邻词中这个词的 $TF_k * IDF_k$ 值作为上下文向量的分量值。其中，停用词被排除在共现值计算之外[孙广范等，2007]。

⑥ 计算向量的相似度：同时遍历 C2 和 E2，从两者中各取一个词形成待计算相似度的词对，对此词对中的两个词对应的两个上下文向量计算向量相似度。采用 Jaccard 相似度计算方法对两个上下文向量进行相似度计算：

$$Jaccard(V,W)=\frac{\sum_k v_k w_k}{\sum_k v_k^2+\sum_l w_l^2-\sum_m v_m w_m}$$

上述公式中，V 和 W 分别代表中英文上下文向量，v_k 和 w_l 是其分量。

⑦ 翻译等价对抽取：将观察窗口长度设为 5，从中文语料出发计算出候选翻译等价对中相似度值大于 0.5 的结果中翻译正确率为 69%。为了提高正确率，再从英语语料出发计算获得与英文对应的中文候选等价对，并将这些翻译等价对与中英的翻译等价对候选集合求交集，最后得到的正确率达到 78%。这验证了双向计算翻译等价对候选集、然后求交集的做法可以提高翻译等价对计算正确率的观点。近年来，分别从可比语料库提取信息技术和体育领域术语 45,000 余条和 38,000 余条，扩大了系统的覆盖面。

3.3 从三元组可比语料库提取本族英语多词表达

由于平行语料库本身存在的翻译语言对本族语言固有的扭斜，使得从平行语料库提取数据的准确度受到一定程度的影响，在汉英平行语料库中扭斜问题更为突出，例如受翻译腔影响把“电子政务建设”、“二手资料”和“重要意义”等错误地翻译成“e-government construction”、“second-hand data”和“important significance”，严重影响了从平行语料库提取多词表达的质量（见表 1）。

建造由标准中文、本族英语和中式英语组成的百万句对级三元组可比语料库，利用统计方法分析关键词簇在词语层面上的过使用和欠使用的语言现象，使用对数似然值（LL）定量分析关键词簇的差异显著性。依据对数似然值的变化差异（见表 1），可以发现中式英语与本族英语多词表达的区别特征，提取本族英语多词表达改进机器翻译系统性能。该方法的对数似然值计算方式如下：

假设 X 为考察的关键词簇，a 和 b 分别为中式英语和本族英语语料库中出现 X 的次数，c 和 d 为中式英语和本族英语语料库中所有关键词簇的数目[胡小鹏等，2014]。

那么对数似然值为：

$$LL=-2\ln\lambda=2\sum_i O_i\ln\left(\frac{O_i}{E_i}\right)$$

O_i 为观察值， E_i 为期望值，其计算方法如下：

$$E_i=\frac{N_i\sum_i O_i}{\sum_i N_i}$$

中式英语语料库中所有关键词簇的数目为 N1 = c，本族英语语料库中所有关键词簇的数目为 N2 = d，那么中式英语和本族英语中关键词簇的期望为：

E1=c*(a+b)/(c+d)；

E2=d*(a+b)/(c+d)；

依据上述公式得到的 E1 和 E2，我们可以求得 LL 值：

LL=2*((a*log(a/E1))+(b*log(b/E2)))

如表 1 示例给出的，对数似然值最大的关键词簇排在列表的顶端，表明该词簇在本族英语和中式英语之间频次分布差异比较大。到目前为止，已提取本族语言多词表达 56,000 余条，明显提高了系统的准确度和流利度。

表 1 中式英语与本族英语多词表达差异显著性剖析结果

关键词	中式英语语料库	本族英语语料库	过使用(+)欠使用(-)	对数似然值	中文表达
	归一化频率	归一化频率			
network bubble	515	36	+	497.82	网络泡沫
Dot-com bubble	16	372	-	404.52	
e-government construction	126	3	+	150.34	电子政务建设
e-government development	9	120	-	113.55	
second-hand data	62	0	+	85.95	二手资料
indirect data	2	58	-	65.64	
Olympic five rings	20	0	+	27.72	奥运五环
The Olympic rings	0	24	-	33.27	
middle-sized	35	1	+	40.77	中等大小
medium-sized	4	30	-	22.50	
important significance	35	3	+	31.69	重要意义
great significance	6	42	-	30.37	

3.4 从平行语料库提取翻译模板

翻译模板是由常量和变量组成的序列。模板中含有变量，因而比词典具有更广的适应性。模板中的常量（包括由多字短语组成的常量）使得其译文比经过语法分析转换生成的译文更为流利自然。因此它是规则主导型混合机器翻译的一项重要组成部分。

3.4.1 模板库的建立方法

基于模板的机器翻译最重要的前提是获取高质量的翻译模板。模板的提取如图 3 所示。

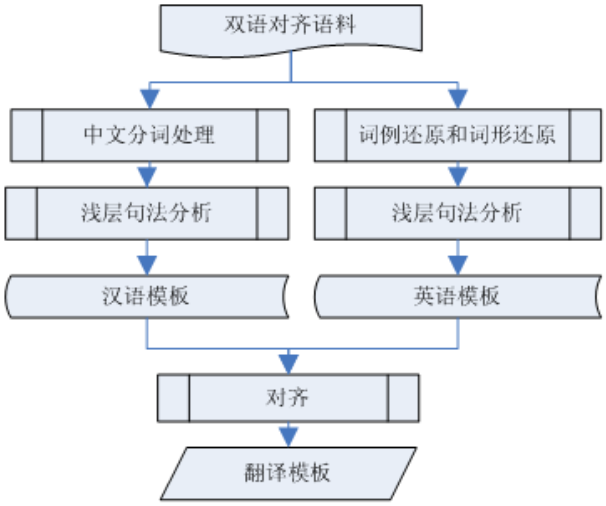


图 3 翻译模板提取

通过对双语句对齐语料的原文和译文句子分别进行自动预处理、分词、词性标注和浅层

句法分析，这样，就得到原文句和对译句的句法树。以该句法树为对象，得到各自独立的语言模板，根据原文和译文的匹配对，进行双向模板之间的对齐，获取模板之间的对应关系，构成翻译模板候选对。然后对该翻译模板候选对进行判别，我们采用多特征的判别模型，设 S 是 N 组候选翻译模板对的集合，S_{ij} 表示第 i 组第 j 个候选翻译模板，其特征表示为 $f_k(S_{ij})$。S_{ij} 的得分如以下公式所示：

$$Score(S_{i,j}) = \sum_{k=1}^{K} f_k(S_{i,j}) \times \lambda_k$$

λ_k 是 $f_k(S_{i,j})$ 对应的权值。K 是总的特征个数。判别模型计算各组中每个候选翻译模板的得分，然后取得分最高的 15%候选翻译模板作为待考证的翻译模板，并加入机器翻译系统中。

3.4.2 翻译模板评估

对翻译模板进行了以下两种不同的评估：

① 模板提取准确度评估

以中欧信息社会语料作为测试语料，其中对 30 万句对进行词处理、浅层分析和翻译模板库提取。平均准确率达到 78 %。

② 对翻译系统性能改进评估

从上述领域测试语料库中随机选取 1,000 句评估翻译模板对系统性能的改进。

评估结果显示英汉双向混合翻译系统的平均准确度均达到 80%以上。

4 系统性能评价与主要应用

4.1 系统性能评价

近几年来，通过利用上述 4 种统计方法在基于规则的核心系统中不断融合从平行和可比语料库提取的术语、多词表达和翻译模板资源，显著增强了混合机器翻译系统的领域适应性和覆盖面。最近使用用户提供的 100 万句对的信息技术和体育领域训练用平行语料库对该混合机器翻译系统进行了测试。测试结果表明，与核心系统相比，翻译准确度提高 15-20 个百分点，达到 80%以上。这里仅以英汉机器翻译为例，通过与谷歌在线翻译系统对比，进一步验证语言学知识模型与数据驱动的统计方法相结合的规则主导型混合机器翻译系统在处理长距离调序，生成地道的句法结构和结构语义方面明显优于统计机器翻译系统。

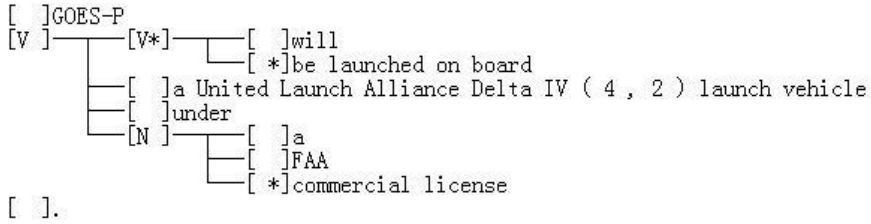
① 系统输入源语句：

GOES-P will be launched on board a United Launch Alliance Delta IV (4, 2) launch vehicle under a FAA commercial license.

② 系统调用半自动生成的机读翻译模板：

[04701] {SUB} {G} be launched on board {OBJ} {I} {N}-->%NODE[%4,%5]，%3%2 搭载%1 发射升空 //weight=-107

③ 混合机器翻译系统生成的句法树：



④ 混合机器翻译系统生成的翻译结果（2014 年 7 月测试）：

在美国联邦航空局商业许可下，联合发射同盟的一枚德尔塔-4（4,2）运载火箭将搭载 GOES-P 卫星发射升空。

⑤ 谷歌在线翻译生成的结果（2014 年 7 月测试）；

GOES-P 将在船上联合发射联盟德尔塔 IV（4，2）根据美国联邦航空局商业许可运载火箭发射。

4.2 系统主要应用

鉴于规则主导型混合机器翻译系统的性能优势，在一些国家重点项目中得到推广使用。自 2007 年该系统成功中标北京奥运会语言自动翻译工程以来，相继在中国-欧盟信息社会项目和公安部天网工程等项目中部署了该系统。此外，CA、Symantec 等跨国 IT 公司已把该系统嵌入本地化流程。

5 下一步工作

今后，规则主导型混合机器翻译研发将重点加强以下四个方面语言学模型与统计方法的融合研究。首先，要继续改进和发展可比语料库建造和使用技术，提高双语数据资源提取准确度和召回率；其次，实现利用自动评价技术发现和定位翻译系统中的语法错误，最大限度提高系统翻译质量和开发效率；第三，研发语言学模型与统计方法相结合的混合术语提取技术，实现“一键式”机器翻译领域适应性解决方案，取代传统的机器翻译定制化开发方式；最后，我们将继续深化与 Lancaster 大学的合作，开展利用多重相关性度量（MAM）方法提高多词表达提取的精确度和召回率以及利用 USAS 语义分析系统增强机器翻译词义消歧（WSD）能力的研究。

本文获国家自然科学基金项目（No.61172101，No.61172102）支持。

参考文献

Reinhard, Rapp et al. 2014. Introduction to The Third Workshop on Hybrid Approaches to Translation. Proceedings of the 3rd Workshop on Hybrid Approaches to Translation (HyTra) @ EACL 2014. Pages iii.

Lori, Thicke. 2013. A Technology Agnostic Approach to Machine Translation: Machine Translation Summit XIV. Proceedings of the XIV Machine Translation Summit. Pages 309-311.

Marta, R. Costa-jussà, Reinhard Rapp. 2013. Workshop on Hybrid Approaches to Translation: Overview and Development. Proceedings of the Second Workshop on Hybrid Approaches to Translation. Pages 1-6.

Reinhard, Rapp et al. 2014. Extracting Multiword Translations from Aligned Comparable Documents. Proceedings of the 3rd Workshop on Hybrid Approaches to Translation (HyTra) @ EACL 2014. Pages 87-95.

ACCURAT. 2012. Analysis and Evaluation of Comparable Corpora for Under Resourced Areas of Machine Translation. In Proceedings of the 16th EAMT Conference. Trento, Italy. Pages 205.

Béatrice, Daille. 2012. Building bilingual terminologies from comparable corpora: The TTC TermSuite. Proceedings of the 5th Workshop on Building and Using Comparable Corpora. Pages 29-32.

Bogdan, Babych et al. 2009. Automated error analysis for multiword expressions: using BLEU-type scores for automatic discovery of potential translation errors. Linguistica Antverpiensia, New Series (8/2009).

Pierre, Isabelle. 2007. Domain adaptation of MT systems through automatic post-editing. In Proceedings of the Machine Translation Summit XI. Pages 255-261.

Michel, Simard, Cyril, Goutte et al. 2007. Statistical phrase-based post-editing. In NAACL-HLT. Pages 508-515.

Loic, Dugast et al. 2009. Statistical post editing and dictionary extraction: Systran/Edinburgh submissions for ACL-WMT2009. In Proceedings of the 4th Workshop on Statistical Machine Translation. Pages 110-114.

Antonio, Lagarda. 2009. Statistical Post-Editing of a Rule-Based Machine Translation System. Proceedings of NAACL HLT 2009: Short Papers. Pages 217-220.

Terumasa, EHARA. 2010. Machine translation for patent documents combining rule-based translation and statistical post-editing. Proceedings of NTCIR-8 Workshop Meeting, Tokyo, Japan. Pages 384-386.

Petra, Wolf et al. 2011. From Statistical Term Extraction to Hybrid Machine Translation. Proceedings of the 15th Conference of the European Association for Machine Translation. Pages 225-232.

Petra, Wolf. 2013. Hybrid Domain Adaptation for a Rule Based MT System. Proceedings of the XIV Machine Translation Summit. Pages 321-328.

Kurt, Eberle，Bogdan Babych et al. 2012. Design of a hybrid high quality machine translation system. Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, Pages 101-112.

冯志伟. 一个新兴的术语学科--计算术语学. 术语标准化与信息技术. 2008，52(4)：4-9.

孙广范，宋金平，袁 琦. 中英可比语料库中翻译等价对抽取方法研究. 计算机工程与应用. 2007，43(32)：44-46.

胡小鹏，袁琦，耿鑫辉. 构建和剖析中英三元组可比语料库. 计算机工程与应用. 2014，50(13)：153-157.