文章编号：1003-0077(2015)02-0001-09

多策略机器翻译研究综述

李业刚[1,2]，黄河燕[1]，史树敏[1]，冯 冲[1]，苏 超[1]

（1. 北京理工大学 计算机学院 北京市海量语言信息处理与云计算应用工程技术研究中心，北京 100081；
2. 山东理工大学 计算机科学与技术学院，山东 淄博 255049）

摘 要：该文全面综述和分析了多策略机器翻译的研究。根据所采用策略方式的差异，我们将多策略机器翻译分为系统级策略融合和模块级策略融合。在分别介绍了不同的翻译方法后，着重介绍了系统级策略融合和模块级策略融合各自具有代表性的研究工作。最后，对多策略机器翻译的研究进行了展望。

关键词：机器翻译；多策略机器翻译；融合机器翻译；混合机器翻译；多引擎机器翻译

中图分类号：TP391 **文献标识码**：A

A Survey of Multi-Strategy Machine Translation

LI Yegang[1,2], HUANG Heyan[1], SHI Shumin[1], FENG Chong[1], SU Chao[1]

(1. Beijing Engineering Applications Research Center of High Volume Language Information Processing and Cloud Computing, School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081, China;
2. Department of Computer Science and Technology, Shandong University of Technology, Zibo Shandong 255049, China)

Abstract: This paper presents an overview of multi-strategy machine translation (MT). According to different level of combination the approaches to multi-strategy MT are classified into system-level combination and module-level combination. The representative method for each combination type are discussed in this paper, and the future development prospects of multi-strategy MT are also discussed.

Key words: machine translation; multi-strategy MT; system combination for MT; hybrid-MT; multi-engine MT

1 引言

机器翻译(Machine Translation, MT)是利用计算机实现从一种自然语言到另一种自然语言的自动翻译技术。机器翻译方法目前公认可以分为基于规则的机器翻译(Rule-Based MT, RBMT)和基于语料库的数据驱动的机器翻译(Corpus-Based MT, CBMT)。RBMT 由词典、规则库以及各类知识库构成知识源；CBMT 以语料应用为核心，进一步分为统计机器翻译(Statistics MT, SMT)、基于实例的机器翻译(Example-Based MT, EBMT)和翻译记忆 (Translation Memory, TM)。RBMT 主要从语言现象着手，侧重描述语言构成规律，对语言规律有良好的概括以及描述能力。SMT 主要从数学角度，侧重统计建模，因而具备良好的数学模型、鲁棒性以及自学习能力。EBMT 是从机器学习的角度，侧重待翻译实例的抽象程度，EBMT 和 TM 对有较高相似度句子的翻译颇有价值。

上述单一的机器翻译方法，很难说哪一种在翻译效果上具有压倒性的绝对优势，它们都存在一些自身难以克服的弊端。同时，它们互不排斥，各自着眼于不同角度，侧重翻译问题的不同侧面。多层次的语言成分构成、严谨的统计数学模型以及丰富的翻译实例，三者是可以共融共生的。在这样的背景下，博采众长，融合多种翻译方法的多策略机器翻译思想应运而生，并成为当前机器翻译研究的热点之一。

多策略机器翻译(Multi-Strategy MT，MSMT)的任务是扬长避短，协调不同翻译策略，融合多种翻

收稿日期：2014-05-19 **定稿日期**：2014-07-16

基金项目：国家重点基础研究发展计划(973)(2013CB329303)；国家自然科学基金(61132009)；国家自然科学基金(61202244)

译方法,从而进一步提升翻译性能。MSMT中的策略,可以是某一种翻译方法,也可以是一种翻译方法中的某个模块。MSMT目前研究呈现出多样化的趋势,可能是两个或者更多的翻译方法的系统级融合,也可能是属于不同翻译方法模块的模块级融合,或者是两者的结合。

我们在前贤们的研究基础上,详尽地介绍了各种不同的翻译策略及其融合方式。文中首先简单介绍了四种翻译方法的基本原理及其优缺点和主要研究热点;然后按照融合方式的不同,分别介绍了系统级的策略融合以及模块级的策略融合;接着介绍了一个典型的MSMT系统;最后对全文进行了总结并对MSMT的发展进行了展望。

2 RBMT

2.1 RBMT概述

RBMT用人工撰写的规则来描述语言规律,首先对待翻译的源语言句子进行分析或理解,对其意义进行表示,然后在某一平面进行语言的转换,最终结合目标语言结构规则生成与源语言等价的目标语言句子。在RBMT中,为了控制规则冲突,保证良好的规则可扩展性,规则往往具有层次性和模块性。因此,RBMT规则系统的组织不仅仅要解决方法论问题,还要从软件工程以及知识工程的层面统筹设计。不同RBMT系统的技术差别主要体现在转换平面上,如词法、句法语义以及语用层面等。

2.2 RBMT的研究

RBMT目前的研究主要集中在基于语义层面的转换,以及多语言翻译特别是少数民族语言翻译。

文献[1]提出了基于语义单元理论的机器翻译方法原理,把自然语言间的翻译看作同一语义在两种自然语言上的不同表示之间的转换,首先,在源语言端进行语义分析,得到句义表达式,然后代入目标语言语义单元表示,生成目标语言句子。文献[2]研究了量词选择,英语介词的语义消歧以及汉英时态转换。文献[3]阐述了一个良构的自然语言句子生成系统。

机器翻译中的本体是对知识的形式化,是语义表达的依托和语义推理的依据,是独立于特定语言的概念库,它可以为词典、语义表示提供语义概念,把语义概念组织为概念层次网络,形成语义空间[4]。计算机可以通过搜索从语义空间中获取有关概念的信息,进行语义计算及推理,从而提高MT系统的语义处理能力,解决在限定翻译领域的一些实际问题。文献[5]对机器翻译专业领域分类系统、专业词典向专业领域分类系统的映射以及国际标准分类ICS标准向专业领域分类系统的映射等问题进行了研究。基于已经构建的领域本体MPO,文献[6]提出一种本体知识规则与统计方法相结合的领域命名实体识别方法,通过本体化的实例,获取构成实体的词性规则模板,进而结合机器学习,识别限定领域命名实体。文献[7-8]提出一种基于词典中注释信息的词汇领域标注方法,利用通用词典中词汇的注释信息给词语标注领域,扩充了现有领域词典的规模。

文献[9]提出层次语义类型树(Semantic Category Tree,SCT)模型,并应用在汉英机器翻译中,实现汉英SCT层面转换,它们为概念层次网络(Hierarchical Network Concepts,HNC)概念体系的3 000多个概念基元建立了概念基元知识库,包括概念基元符号、概念基元延伸节点的表示以及概念关联表示式等,用概念延伸结构表示代替了本体的上下位表示方法,不同的延伸结构代表了不同的语义扩展,同时,概念之间的关系用概念关联表示式描述,并利用概念和词语的绑定来增强词语的聚集性。

另外,在少数民族语言方面,文献[10]描述了统一标准、接口的多民族语言本体知识库的创建思路。文献[11]建立了蒙古语的语义知识库。文献[12]阐述了维语的框架语义描述体系。

2.3 RBMT的优势及存在的问题

RBMT历经几十年的不断发展,不断融入人工智能的最新成果,日趋完善。RBMT直观地表达语言学知识,良好地概括和描述语言规律,详尽的规则能够准确、直观地描述语言的语法、语义构成,多层次的规则便于进行深层理解和复杂结构处理,对不同句子实施不同平面转换,有效解决长距离依赖问题。真正为用户所使用的专业机器翻译产品大多都是基于规则的系统。

因为规则库是众多的语言学家手工构建的,所以一致性很难保障,当规则库达到一定规模后,进一步扩充规则非常困难。由于语言现象庞杂,现有的理论方法和语言规则都无法有效地表达所有语言现象,趋于无限的语言现象和枚举的规则系统之间的矛盾是RBMT的局限性,这也最终影响了RBMT在开放领域中的适应性。

3 SMT

3.1 SMT 概述

SMT 把翻译看作概率问题，认为任意一个目标语言句子都在一定概率上是任意一个源语言句子的译文，SMT 的目标就是找到概率最大的那个目标语言句子。SMT 的首要任务是模型问题，就是为机器翻译建立合适的概率模型，确定源语言句子到目标语言句子的翻译概率的计算方法，并在此基础上，定义要估计的参数，设计估计的算法。SMT 奠基性的工作是文献[13]提出的信源信道模型，对后继的 SMT 研究产生了深远的影响，噪声信道模型如式(1)所示。

$$\begin{aligned} e^* &= \arg\max P(e \mid f) \\ &= \arg\max P(e)P(f \mid e) \end{aligned} \tag{1}$$

信道模型包括三个基本组件：翻译模型P(f|e)、语言模型 P(e)以及解码。翻译模型计算目标语言句子和源语言句子的翻译概率；语言模型对生成的目标语言句子进行评估，保证其流畅性；解码是在已知模型以及相关参数的基础上，对于任何一个源语言句子，查找翻译概率最大的目标语言句子。

语言模型[14](language model, LM)是 SMT 系统中的重要模块，它被用来衡量翻译系统输出句子的流畅程度，给定一个词汇序列 $\omega_1,\omega_2,\cdots,\omega_n$，n 元语言模型的计算如式 2 所示。

$$p(\omega_1,\omega_2,\cdots,\omega_n) = \prod_{i=1}^{N} p(\omega_i \mid \omega_{i-n+1},\cdots,\omega_{i-1}) \tag{2}$$

它有一个重要假设，即当前词汇 ω_i 出现的概率仅与前 n-1 个词汇 $\omega_{i-n+1},\cdots,\omega_{i-1}$ 相关，而与其他词汇无关。

文献[15-16]将对数—线性模型(log-linear)引入 SMT，提出了基于短语的统计机器翻译(Phrase-Based MT, PBSMT)，该模型对 P(e|f)进行建模，能够整合各种不同的特征(feature)，并允许自动调节特征的权重，将连续的多词作为短语，整体翻译，扩大了翻译的粒度，容易处理局部上下文依赖关系，能够较好地翻译习语和常用搭配。这项工作对 SMT 的发展影响重大，几乎现在的 SMT 全部是采用对数线性模型框架。其数学表达形式如式(3)所示。

$$\begin{aligned} \hat{e} &= \arg\max_e P(e \mid f) \\ &= \arg\max_e \frac{\exp\left(\sum_{m=1}^{M} \lambda_m h_m(e,f)\right)}{\sum_{e'} \exp\left(\sum_{m=1}^{M} \lambda_m h_m(e',f)\right)} \\ &= \arg\max_e \left(\sum_{m=1}^{M} \lambda_m h_m(e,f)\right) \end{aligned} \tag{3}$$

3.2 SMT 的研究

目前，SMT 的研究[17-19]集中在将句法知识引入到翻译框架中，利用句法知识来限制翻译路径，约束目标词和短语的活动范围。典型的研究有吴德凯[20]和 Chiang[21]的基于形式化句法的翻译模型以及南加州大学信息科学研究所提出的树—串翻译模型[22]。相比传统的基于短语的翻译模型，层次短语翻译模型能够处理非连续短语，并具有一定的泛化能力，且不受句法分析的制约。基于语言学语法的统计机器翻译则包含了丰富的语言学知识。

尽管基于句法的 SMT 具有一定的长距离调序的能力，但是纯粹的基于句法的 SMT 受限于双语句法结构的不一致性、生成规则中的终结符过分泛化以及生成规则的规模过于庞大等因素，翻译质量并没有显著提高。如何在 SMT 中更有效地融入句法知识，既保证对句法知识的容错能力，又能够解释不同语言之间的差异，还需要进一步的深入研究。

另外，目前研究主要是尝试在句法层面融入语言学知识。如何选择一种可计算、表达能力强的表示形式，如何选择一种有较强的数据学习能力的合适模型，把更深层次的语言学知识，比如语义知识和篇章上下文知识，有效融入 SMT 框架，也需要进一步的研究。

随着各种资源越来越丰富以及算法的日趋复杂，SMT 的计算量也越来越大。Google 之所以在机器翻译领域占据领先地位，也是源于其能力强大的分布式计算。因此，结合分布式计算与机器翻译，将机器翻译相关计算进行并行化处理也将是 SMT 的研究热点。

3.3 SMT 优势及目前存在的问题

SMT 由于具有良好的数学模型、自学习能力和鲁棒性等优点，从而备受研究者的钟爱，迅速被开放领域的互联网机器翻译所采纳，成为目前非限定领域机器翻译中表现最佳的一种翻译方法。

SMT 依赖于大规模的双语语料，依靠统计进行歧义的消解以及译文的选择。翻译模型以及语言模

型的参数估计的准确性都直接依赖于语料的规模，翻译效果最终取决于概率模型和语料库的覆盖能力。因此，对于语料匮乏的语言之间的翻译，比如我国的少数民族语言，能力有限。SMT还面临数据稀疏问题。即便是在超大规模的语料库中，也会存在相当一部分的低频词，低频词的统计信息往往不够准确，这些不准确的统计最终会影响SMT的翻译性能。

单纯依赖统计量的SMT难于反映语言真实的内部规律，简单的统计量也很难解释差异较大语言之间的复杂结构对应关系，这就造成翻译结果虽然“词词相对”，却不具备可读性，晦涩难懂。

4 EBMT

4.1 EBMT概述

EBMT是以翻译实例为出发点基于类比原理的机器翻译方法。EBMT把源语言句子分解为片段，通过类比找到这些片段对应的目标语言的片段，经过对目标语言片段的适当重组，形成句子翻译结果。EBMT主要的知识源是双语对照的实例库和义类词典等，其核心问题是通过最大限度的统计，得出双语对照的实例库。不同的EBMT系统之间的主要区别在于相异的双语语料库结构以及翻译模板以及翻译模型的生成技术不同。

4.2 EBMT的研究

句子之间的相似可以表现在语义、结构、目标特征和个体特征等不同方面。根据类比推理，最优匹配最好要同时满足前述的约束。然而，语言的无穷性将会导致模板库趋于无穷大。为了增强模版的覆盖能力，在构造模板时，可以对实例进行适当的泛化(Generation)，把句中一些不影响整体结构和总体表达的可替换的成分抽象化，从而降低输入的维数，提高句子的匹配率。基于模板的机器翻译方法(Template-based MT/Pattern-based MT，TBMT/PBMT)是EBMT翻译方法的扩展，是EBMT中的一种典型翻译方法。

实例的泛化程度可高可低，既可以是将双语实例中的特殊语言成分(比如命名实体等)用类标表示[23]；也可以是将句子中相同部分表示为变量，泛化后的句子模板是比规则更具体比实例更抽象介于规则和实例之间的知识粒度，模板的粒度将直接影响到匹配的效果。利用语法或者语义概念层次结构的源语言句子的相似度的计算以及限制翻译模板的变量是翻译模板研究的趋势之一[24-26]。

4.3 EBMT优势及存在的问题

EBMT系统能够利用翻译实例中隐含的结构信息对译文中的词进行约束，一般不对源语言进行深层次分析，对于实例库中的已有句子，可以直接高质量翻译，对实例库中存在与实例比较相似的句子，可以通过类比推理，并对翻译结果进行少量的修改后，近似翻译。EBMT还可以同时给出翻译结果的置信度，这也是EBMT在系统融合中备受欢迎的一个重要原因。

EBMT需要对语言的互译片段建立映射，即短语甚至词汇一级的双语对齐。短语对齐往往存在歧义，这将影响译文的质量。不进行语言深分析的EBMT系统，缺乏句子的深层结构信息，翻译碎片组合比较困难，生成的译文信息往往有所匮乏。而基于深层次分析技术的EBMT系统，因为各种语言分析器训练语料的不平衡，在不同应用领域上的性能差别非常大。

EBMT把训练过程放在了解码阶段，翻译实时性会受到较大的影响，其受限于大规模实例语料库中相似实例的检索速度。

对于可检索到相似实例的源句子，EBMT能够生成高质量的译文。因此，实例的覆盖率是EBMT系统的重要因素，但受限于语料库规模，EBMT很难达到较高的匹配率，往往只有在限定领域和专业领域，翻译效果才能达到使用要求。因此，单纯采用RBMT的系统较少，一般都把它作为多翻译引擎中的一个。

5 TM

TM是利用已有的源语言资源和对应的目标语言资源，建立起一个或多个翻译记忆库。在翻译过程中，TM系统自动搜索翻译记忆库中相同或相似的翻译资源(如句子、篇章)，作为参考译文呈现给用户。用户可以选择接受参考译文，也可以在译文基础上进行修改，得到最终的译文。用户修改过的译文和对应的源文会自动存入记忆库，供下次使用。TM系统的性能与翻译资料的重复性有很大的关系，重复性内容越多，翻译效果就越好。

TM所面对的用户通常是领域的“专家”，这与

EBMT 不同，EBMT 翻译的结果由系统决定，用户只需要懂目标语言即可。从这点来说，TM 不是纯粹的机器翻译方法而是属于辅助机器翻译。但是 TM 与 EBMT 存在许多相似的地方，例如，对已有翻译实例的重用，翻译实例的存储，相似翻译实例的检索等。所以，实际研究中，研究者们经常忽略它们的不同，把 TM 也看作一种机器翻译方法。

6　系统级策略融合

针对单一的机器翻译方法本身及发展中存在的问题[27-28]，系统级策略融合（也称作融合机器翻译（System Combination for MT）、混合机器翻译（Hybrid MT）或者多引擎机器翻译（Multi-Engine MT）），致力于在后处理或是翻译过程中，扬长避短，融合多个机器翻译引擎的有用信息，得到更好地译文。按照融合的阶段可分为后处理级系统融合和模型间系统融合。国内机器翻译评测会议（CWMT）率先从 2008 年开展系统融合单独评测，国际机器翻译评测 NIST 也从 2009 年开始将系统融合作为单独的项目进行评测，这也从另一方面说明系统融合技术的重要性。

6.1　后处理系统融合

在后处理系统融合中，融合可以在句子、短语或者词粒度上独立进行[29]，也可以结合起来进行。句子粒度的系统融合可以是并列式系统融合，也可以是递进式系统融合。并列式系统融合平等的对待所有的融合系统，针对同一个源语言句子，使用单机器翻译引擎所使用的特征之外的特征，从合并后多个系统的翻译结果的 N-best 列表中找出翻译质量最高的结果，实际上是一种句子重排序，目前的研究主要集中在对融合策略的探索。基于最小贝叶斯风险（Minimum Bayes-Risk Decoding，MBR）[30]的系统融合方法是从多个系统的翻译结果的 N-best 列表中选择期望损失最小的，如式（4）所示。

$$\hat{e} = \arg\min_{e' \in E_h} \sum_{\substack{ref \in E_h \\ ref \neq e'}} L(e', ref) * p(ref \mid f) \quad (4)$$

在这里 E_h 代表由多个机器翻译系统结果组成的 N-best 列表；ref 表示参考译文。$L(e', ref)$ 表示损失函数，它的值越小，对应翻译结果 e′的质量越高；P(ref|f)代表翻译后验概率，系统融合的输入来源较多，不同系统给出的后验概率不具备可比性，RBMT 系统则无法给出后验概率，因此在使用中往往设置的后验概率是相同的。

通用线性模型[31]（generalized linear model）把翻译假设所对应的翻译的置信度取对数，与高阶语言模型（例如，5 阶）得分以及长度惩罚线性加权，作为评分准则，如式（5）所示。

$$L(e_i) = \log(p(e_i)) + \nu LM^{5-gramm}(e_i) + \mu \mid e_i \mid \quad (5)$$

其中，$p(e_i)$ 为翻译假设 e_i 对应的翻译置信度，ν 和 μ 分别为五元语言模型和长度惩罚 $\mid e_i \mid$ 对应的特征权重，这些权重可以在开发集上进行优化得到。

文献[32]提出了一种基于机器学习的翻译推荐策略，对于 MT 系统的输出和 TM 系统的参考翻译，通过分类器挑选出更适合后编辑的译文，呈现给用户，进行人工后编辑，该方法把判断哪一个输出结果适合后编辑问题看作是一个分类，使用翻译编辑率（Translation Edit Rate，简称 TER）[33]来自动评价后编辑的工作量。后编辑所需工作量最小的结果，并不一定是 SMT 或 TM 的 Top-1 结果。因此为了更好地利用两个系统的 N-best 结果，文献[34]提出了一个基于重排序的翻译推荐方法：对于 SMT 和 TM 的 N-best 结果，利用支持向量机（Support Vector Machine，SVM）进行重新打分排序，并将新产生的 Top-n 结果，人工进行后编辑。采用的改进的优化函数如式（6）所示。

$$\begin{aligned}
&\min_{\omega,b,\xi} \frac{1}{2} w^T w + C \sum \xi \\
&\forall (d_i, d_j) \in r(s_1): w(\Phi(s_1, d_i) - \Phi(s_1, d_j)) \geqslant 1 - \xi_{i,j,1} \\
&\cdots \\
&\forall (d_i, d_j) \in r(s_n): w(\Phi(s_n, d_i) - \Phi(s_n, d_j)) \geqslant 1 - \xi_{i,j,n} \\
&\xi_{i,j,k} \geqslant 0, 1 \leqslant k \leqslant n
\end{aligned} \quad (6)$$

其中，$\Phi(s_n, d_i)$ 表示给定源语言句子 s_n 对应翻译输出 d_i 的特征向量。诸如此类的基于机器学习的策略融合研究还有文献[35-36]。

文献[37-38]在多 Agent 的日汉机器翻译系统中，采用 TM、EBMT、RBMT 多种机器翻译方法相结合的递进式融合，机器翻译的流程分为三个递进式模块，从基础的 TM 翻译，到需要源语言句法信息的 EBMT，再到最复杂、需要源语言句法、语义分析的基于配价和断段分析的 RBMT，当前一个翻译模块的译文评分达到设立的阈值时，该模块的译文输出作为最终翻译结果，否则进入下一个模块进行更深层的处理。系统取得了较好的翻译效果，其中，

在开放测试中,译文可读性达到了79%。

句子粒度的系统融合不生成新的翻译假设,有效地保护了原来翻译假设中句子的词序以及短语的连续性。但是,融合后输出的译文也不能借鉴其他翻译假设中词或短语粒度的知识,仅仅从句子粒度对翻译假设进行横向比较,因此,它对翻译性能的提高相对较小。递进式策略融合相比并列式策略融合,在翻译系统选择上有约束,要求参与融合的翻译系统间差异比较大,一般选择TM、EBMT这种特色鲜明的翻译引擎。优点是融合性能稳定,翻译质量会有稳步的提高。同时,由于参与融合的翻译引擎不是并列的,翻译效率也比较高,翻译时间约等于$\max(t_i), i=1,2,\cdots n$,而并列式融合系统的翻译时间大于$\sum_{i=1}^{n} t_i$,其中$t_i$表示第i个翻译引擎的翻译时间,n表示参与融合的翻译引擎的个数。相比递进式融合,并行融合方式能融入更多的翻译引擎。

目前国内外后处理系统融合研究热点集中在词粒度的系统融合[29],借鉴语音识别中混淆网络解码[39]的思想,将多个翻译系统输出的翻译假设,利用词对齐方法构建混淆网络(或称为词转换网络),对混淆网络中每一个位置的候选词进行置信度估计,最后进行混淆网络解码。这种融合方法在词的层次重组了输出译文,因此能够充分利用各个翻译假设的词汇粒度的知识,取长补短。混淆网络解码同时也破坏了原来的翻译假设的词序的一致性以及短语连贯性,因此,也会发生融合后的译文不符合语法的情况。

6.2 模型间系统融合

模型间的融合是利用机器学习算法,在更深层次融合两个具有互补性的翻译模型,从而提高翻译性能。

TM或者EBMT引入到SMT,相当于在SMT中间接利用了全局信息,将会改善SMT系统输出,推动SMT在专业翻译领域的应用。文献[40]提出,首先使用EBMT,查找最相似的实例,然后利用句法和词对齐信息,抽取匹配部分的翻译,并利用XML标记法固定匹配部分的翻译,使用SMT系统翻译剩余部分。文献[41]则把XML标记法引入到TM和SMT的融合。实验结果表明,仅当模糊匹配系数高于0.7时,XML标记法才能改善SMT系统翻译性能,否则会导致翻译性能降低。但是,模糊匹配系数低,并不意味着TM中所有的片段都没有价值;模糊匹配系数高,也不意味着TM中所有的片段都有价值,因此,文献[42]提出了决策式XML标记法,使用分类器,代替模糊匹配系数,决定是否使用XML标记法,对于需要进行XML标记的句子,XML标记法保留了匹配短语,剩余部分则SMT进行翻译,对于不需要XML标记的句子通过SMT进行翻译。

上述方法仅在翻译的输出上进行浅层融合,并没有改变SMT模型和解码器,因此性能提升的幅度不大。因此,文献[43]提出了一种在解码层面进行TM和SMT的深层次的融合框架,并引入了模糊匹配区间索引、源语言短语链接状态和目标语言短语匹配状态三种特征集,验证了三种由简到繁的整合式融合模型。当模糊匹配系数大于0.4时,BLEU值和TER值都显著优于单独的SMT和TM系统,BLEU比SMT基线系统提高了3.48个百分点,TER值提高了2.62个百分点。

SMT和EBMT采用相同的词对齐双语语料库,因而可以结合两者的优点,利用EBMT获得实例中蕴含的丰富信息以及相应的翻译结构,利用SMT的各类模型特征定量评价译文的好坏。文献[44-46]提出了混合数据驱动机器翻译模型框架,在法-英翻译任务中的性能优于单一的EBMT和PBSMT。文献[47]在EBMT系统中,加入了类似于PBSMT翻译模型的特征,并综合考虑了上下文特征,进一步提升了翻译性能,BLEU值比基线系统提高接近4个百分点。

同是SMT系统,PBSMT没有考虑句法信息,基于句法的SMT的规则覆盖规模不如PBSMT好,解码中短语的匹配不够灵活,即便同是基于句法的SMT,不同文法的表现力也不同,使用不同文法的SMT也可以进行融合,取长补短。文献[48-50]在机器翻译的解码框架下融合层次短语文法和括号转录文法,考虑了在解码过程生成的侯选翻译相互之间的影响。文献[51]则是在超图的框架下,通过*n*-gram后验概率特征来对PBSMT和基于层次短语的SMT两个翻译模型进行重新搜索,得到翻译结果;文献[52]在超图框架下,通过两个模型的*n*-gram后验概率特征进行线性插值,得到翻译结果,采用了两阶段的最小错误率训练,由于不对生成翻译结果进行重新训练和解码,因而翻译效率比较高,同时翻译性能也优于单个系统。文献[53]把PBSMT的词汇化调序特征和距离惩罚调序特征加入到基于句法树的SMT的解码过程,改善了翻译

性能。

7　模块级策略融合

模块级策略融合是以一种翻译策略为主，在系统中融合属于不同翻译策略的模块。例如，基于规则的分析器、基于统计的词对齐模块、语言模型、后编辑模块等都在不同的翻译系统中得到广泛的应用。相比于全方位的系统级策略融合，模块级策略融合更侧重于融合的灵活性。

文献[54]在基于实例的机器翻译框架下，使用基于规则的分析器对源语言进行分析；使用基于统计的词对齐模块，建立源语言与目标语言间的对应关系；统计语言模型被用来对目标语建模；基于规则的后处理模块被用来做最终的目标语言生成处理。

文献[55]在半结构化的 EBMT 系统中，引入基于统计搭配模型的译文选择方法，估计候选译文中词汇之间的搭配关系，利用编辑距离选择匹配翻译实例，使用源语言统计搭配模型计算词汇间的匹配度，并估计句子中词汇的编辑风险，在英汉翻译中，BLEU 得分比基线 EBMT 系统提高了 4.73～6.48 个百分点。

后编辑(post-editing，PE)是对机器翻译系统输出的译文进行加工和修正。近年来，统计后编辑(Statistical post-editing，SPE)得到了长足的发展[56]，它可以用来改善 RBMT 的翻译性能和领域适应性[57-60]。其中，文献[58]使用基于短语的 SPE 对基于规则的 SYSTRAN 翻译系统进行后编辑，实验证明，即便是在少量训练语料(大于 1M)上训练的 SPE，也可以显著提高基于规则的机器翻译系统 SYSTRAN 的性能，随着训练语料增加，翻译性能得到持续提升，当训练语料增加到 100M 级别时趋于收敛。

结论和展望

随着各种机器翻译方法如火如荼的发展，多策略的机器翻译研究也取得了长足的进步和丰硕的成果。其中，在第七届全国机器翻译研讨会机器翻译评测[27]中，辅以统计后编辑的 RBMT 系统的 BLEU 值(0.238 7)在汉英新闻评测中名列榜首；SMT 和 RBMT 后处理融合系统的 BLEU 值(0.408 3)在英汉科技领域名列第一，这很大程度上促进了机器翻译的整体发展。

尽管如此，现有的多策略翻译仍然达不到令人满意的程度，很多研究尚停留在理论水平，为了进一步推动 MSMT 的发展，笔者认为以下几个方面的研究仍然是值得期待的。

(1) 翻译模型的差异性、翻译模型参数以及 N-best 数量都会影响系统融合的效果。目前，一般选择既有差异性又能够互补，翻译质量相差不太大的翻译模型参与融合，通过在开发集上的组合策略尝试，最终选取最有效的融合方式。相似翻译模型以及质量稍差的翻译模型参与融合是不是完全没有可取之处，以及有没有更好的组合策略代替现有的枚举尝试方式可以进一步研究。

(2) 融合机器翻译目的是获取比单个系统更优的翻译结果，但是目前系统策略融合鲁棒性不充分，存在数据集依赖问题，甚至会出现低于最优的单个系统的翻译性能的情况。保守的策略融合虽然具有较强的鲁棒性，能保证目标翻译的质量，但提高的幅度则比较小。

(3) 不同的后处理融合方法、模型间融合方法以及在更多模块的融入策略各有优势，如何有效地组合，产生有效率和性能兼顾的翻译系统值得期待。

(4) 目前的策略融合的粒度局限于句法层面，如何融入语义等更深层次知识来指导融合，有效地改善翻译质量也值得尝试。

(5) 在尽可能提高融合后机器翻译的性能的同时，也要兼顾融合机器翻译的效率。多个引擎并行融合，需要翻译时间是倍增的。提高翻译效率，快速融合也是 MSMT 系统的一个趋势。

(6) 机器翻译评测本身就是一个人工智能问题，无论基于编辑距离还是 N 元匹配的自动评测都有各自的局限性，用单一的测评方法评价不同的翻译引擎，往往有失公允。MSMT 恰恰融合了多种翻译引擎，因此，制定合适的评价方法，综合考虑句法、语义等层面的信息，公正的评价 MSMT，以评测促发展也是至关重要的。

参考文献

[1] Gao Q S, Hu Y, Li L, et al. Semantic language and multi-language MT approach based on SL[J]. Journal of Computer Science and Technology, 2003, 18(6): 848-852.

[2] 关晓薇. 基于语义语言的机器翻译系统中若干关键问题研究[D]. 大连理工大学博士学位论文, 2009.

[3] 胡玥, 高小宇, 李莉, 等. 自然语言合理句子的生成

系统[J]. 计算机学报，2010,33(3):535-544.
[4] 俞士汶，穗志方，朱学锋. 综合型语言知识库及其前景[J]. 中文信息学报，2011，25(6)：12-20.
[5] 黄河燕,张克亮,张孝飞. 基于本体的专业机器翻译术语词典研究[J]. 中文信息学报,2007,21(1)：17-22.
[6] 史树敏,冯冲,黄河燕，等. 基于本体的汉语领域命名实体识别[J]. 情报学报，2009，6：857.
[7] 朱朝勇. 基于本体的知识库分类研究[D]. 中国科学技术大学博士学位论文，2013.
[8] 朱朝勇,黄河燕,史树敏. 基于词汇注释的层次化领域标注[J]. 中国通信，2012，9(3)：19-27.
[9] 朱小健,晋耀红. 层次语义类型树模型及其在汉英机器翻译中的应用[J]. 中国通信,2012,9(12):80-92.
[10] 赵小兵,邱莉榕,赵铁军. 多民族语言本体知识库构建技术[J]. 中文信息学报,2011,(04):71-74.
[11] 那顺乌日图. 蒙古语语言知识库的建立与应用[J]. 中文信息学报,2011,(06):162-165.
[12] 阿里甫·库尔班,吾买尔江·库尔班,尼加提·阿不都肉苏力. 维吾尔语框架语义知识库的概念设计[J]. 中文信息学报,2010,(04):114-118.
[13] P. Brown, S. Della Pietra, V. Della Pietra, et al. The Mathematics of Machine Translation: Parameter Estimation. Computational Linguistics. 1993,19(2): 263-311
[14] Bahl L R, Jelinek F, Mercer R L. A maximum likelihood approach to continuous speech recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1983 (2): 179-190.
[15] Och F J, Ney H. A comparison of alignment models for statistical machine translation[C]//Proceedings of the 18th conference on Computational linguistics-Volume 2. Association for Computational Linguistics, 2000: 1086-1090.
[16] Och F J, Ney H. Discriminative training and maximum entropy models for statistical machine translation[C]//Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics, 2002: 295-302.
[17] Xiao T, Zhu J. Unsupervised Sub-tree Alignment for Tree-to-Tree Translation[J]. Journal of Artificial Intelligence Research, 2013, 48: 733-782.
[18] 刘群. 基于句法的统计机器翻译模型与方法[J]. 中文信息学报,2011,(06):63-71.
[19] 熊德意,刘群,林守勋. 基于句法的统计机器翻译综述[J]. 中文信息学报,2008,(02):28-39.
[20] Dekai Wu. Stochastic Inversion Transduction Grammars and Bilingual Parsing of Parallel Corpora[J]. Computational Linguistics, 1997, 23:377-404.
[21] Chiang, D. Hierarchical Phrase-Based Translation[J]. Computational Linguistics,2007,33(2): 201-228.
[22] Yamada K, Knight K. A syntax-based statistical translation model[C]//Proceedings of the 39th Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics, 2001: 523-530.
[23] Brown R D. The CMU-EBMT machine translation system[J]. Machine translation, 2011, 25(2): 179-195.
[24] 郝晓燕,刘伟,李茹,刘开瑛. 汉语框架语义知识库及软件描述体系[J]. 中文信息学报,2007,(05):96-100,138
[25] 李茹，王智强，李双红，等. 基于框架语义分析的汉语句子相似度计算[J]. 计算机研究与发展，2013，50(8)：1728-1736.
[26] H. Jiexu Cao Yu and Guan Xiaowei. A Set of Machine Learning Methods for Inducing Translation Templates with Grammar- semantic Type Constraints [J], Information and Control Express LetterS. 2011, 15(3):701-706
[27] 赵红梅，吕雅娟，贲国生，等. 第七届全国机器翻译研讨会机器翻译评测总结[J]. 中文信息学报，2012，26(1)：22-30.
[28] 杜金华，张萌，宗成庆，等. 中国机器翻译研究的机遇与挑战--第八届全国机器翻译研讨会总结与展望[J]. 中文信息学报，2013，27(4)：1-8.
[29] 李茂西,宗成庆. 机器翻译系统融合技术综述[J]. 中文信息学报,2010(4):74-84.
[30] Kumar S, Byrne W J. Minimum Bayes-Risk Decoding for Statistical Machine Translation[C]//Proceedings of the HLT-NAACL. 2004: 169-176.
[31] Rosti A V I, Ayan N F, Xiang B, et al. Combining Outputs from Multiple Machine Translation Systems [C]//Proceedings of the HLT-NAACL. 2007: 228-235.
[32] He Y, Ma Y, van Genabith J, et al. Bridging SMT and TM with translation recommendation[C]//Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2010: 622-630.
[33] Snover M, Dorr B, Schwartz R, et al. A study of translation edit rate with targeted human annotation [C]//Proceedings of association for machine translation in the Americas. 2006: 223-231.
[34] He Y, Ma Y, Way A, et al. Integrating N-best SMT Outputs into a TM System[C]//Proceedings of the 23rd International Conference on Computational Linguistics: Posters. Association for Computational Linguistics, 2010: 374-382.
[35] Federmann C. Multi-Engine Machine Translation as a Lifelong Machine Learning Problem[C]//Proceedings of the 2013 AAAI Spring Symposium Series. 2013.
[36] Federmann C. A machine-learning framework for hy-

brid machine translation[C]//Proceedings of the KI 2012: Advances in Artificial Intelligence. Springer Berlin Heidelberg, 2012: 37-48.

[37] 张捷，陈群秀. 日汉机器翻译系统中的多 Agent 研究[J]. 中文信息学报，2003，17(1)：7-12.

[38] 杜伟，陈群秀. 多策略汉日机器翻译系统中的核心技术研究[J]. 中文信息学报，2008，22(5)：60-66.

[39] SIM K, BYRNE W, GALES M, et al. Consensus network decoding for statistical machine translation system [C]//Proceedings of the IEEE International Conference on Acoustics,Speech, and Signal Processing, 2007: 105-108.

[40] Smith J, Clark S. EBMT for SMT: a new EBMT-SMT hybrid[C]//Proceedings of the 3rd International Workshop on Example-Based Machine Translation. 2009: 3-10.

[41] Koehn P, Senellart J. Convergence of translation memory and statistical machine translation[C]//Proceedings of AMTA Workshop on MT Research and the Translation Industry. 2010: 21-31.

[42] Ma Y, He Y, Way A, et al. Consistent Translation using Discriminative Learning-A Translation Memory-inspired Approach[C]//Proceedings of the ACL. 2011: 1239-1248.

[43] Kun Wang, Chengqing Zong and Keh-Yih Su. Integrating Translation Memory into Phrase-Based Machine Translation during Decoding. To appear in Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL), Sofia, Bulgaria, August 4-9, 2013.

[44] Groves D, Way A. Hybrid example-based SMT: the best of both worlds? [C]//Proceedings of the ACL Workshop on Building and Using Parallel Texts. Association for Computational Linguistics, 2005: 183-190.

[45] Groves D, Way A. Hybrid data-driven models of machine translation[J]. Machine Translation, 2005, 19 (3-4): 301-323.

[46] Groves D. Hybrid data-driven models of machine translation[D]. Dublin City University, 2007.

[47] Liu Z, Wang H, Wu H. Example-based machine translation based on tree-string correspondence and statistical generation[J]. Machine translation, 2006, 20(1): 25-41.

[48] Jiang H, Yang M, Zhao T, et al. A statistical machine translation model based on a synthetic synchronous grammar[C]//Proceedings of the ACL-IJCNLP 2009 Conference Short Papers. Association for Computational Linguistics, 2009: 125-128.

[49] Duan N, Li M, Zhang D, et al. Mixture model-based minimum bayes risk decoding using multiple machine translation systems[C]//Proceedings of the 23rd International Conference on Computational Linguistics. Association for Computational Linguistics, 2010: 313-321.

[50] Xiao T, Zhu J, Zhu M, et al. Boosting-based system combination for machine translation[C]//Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2010: 739-748.

[51] DeNero J, Kumar S, Chelba C, et al. Model combination for machine translation [C]//Proceedings of the Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, 2010: 975-983.

[52] Duan N, Li M, Zhang D, et al. Mixture model-based minimum bayes risk decoding using multiple machine translation systems[C]//Proceedings of the 23rd International Conference on Computational Linguistics. Association for Computational Linguistics, 2010: 313-321.

[53] Nguyen T L, Vogel S, Tower T, et al. Integrating Phrase-based Reordering Features into a Chart-based Decoder for Machine Translation[C]//Proceedings of ACL. 2013.

[54] 王海峰，吴华，刘占一. 互联网机器翻译[J]. 中文信息学报，2011，(06)：72-80.

[55] 刘占一，李生，刘挺，等. 利用统计搭配模型改进基于实例的机器翻译[J]. 软件学报，2012，23(6)：1472-1485.

[56] Rosa R, Marecek D, Tamchyna A. Deepfix: Statistical Post-editing of Statistical Machine Translation Using Deep Syntactic Analysis[J]. ACL 2013, 2013: 172.

[57] Dugast L, Senellart J, Koehn P. Statistical post-editing on SYSTRAN's rule-based translation system [C]//Proceedings of the Second Workshop on Statistical Machine Translation. Association for Computational Linguistics, 2007: 220-223.

[58] Michel Simard, Cyril Goutte, and Pierre Isabelle. 2007. Statistical phrase-based post-editing[C]//Proceedings of NAACL HLT 2007, pages 508-515. Rochester, NY.

[59] Michel Simard, Pierre Isabelle, and Cyrill Goutte. 2007. Domain adaptation of MT systems through automatic post-editing [C]//Proceedings of the MT Summit XI, pages 225-261, Copenhagen, Denmark.

[60] Béchara H, Rubino R, He Y, et al. An Evaluation of Statistical Post-Editing Systems Applied to RBMT and SMT Systems[C]//Proceedings of the COLING. 2012: 215-230.

(下转第 23 页)

Technology，2004. 55(14)：1270-1281.

[40] Berberich K.，Vazirgiannis M.，Weikum G. Time-Aware Authority Ranking[J]. Internet Mathematics，2005. 2(3)：301-332.

[41] Wan J.，Bai S. An improvement of PageRank algorithm based on the time-activity-curve[C]//The 2009 IEEE International Conference on Granular Computing. 2009：549-552.

[42] Salton G. The SMART Retrieval System：Experiments in Automatic Document Processing[M]. Upper Saddle River. Prentice-Hall，Inc. 1971.

[43] Lavrenko V.，Croft W. B. Relevance based language models[C]//Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval. Louisiana，USA. 2001：120-127.

[44] Amodeo G.，Amati G.，Gambosi G. On relevance，time and query expansion[C]//Proceedings of the 20th ACM international conference on Information and knowledge management. Scotland，UK. 2011：1973-1976.

[45] Peetz M.-H.，Meij E.，Rijke M. d.，et al. Adaptive temporal query modeling[C]//Proceedings of the 34th European conference on Advances in Information Retrieval. Barcelona，Spain. 2012：455-458.

[46] Keikha M.，Gerani S.，Crestani F. Time-based relevance models[C]//Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval. Beijing，China. 2011：1087-1088.

[47] Miyanishi T.，Seki K.，Uehara K.，Combining Recency and Topic-Dependent Temporal Variation for Microblog Search[C]//Advances in Information Retrieval. Berlin Heidelberg. 2013：331-343.

[48] Whiting S.，Klampanos I. A.，Jose J. M. Temporal pseudo-relevance feedback in microblog retrieval[C]//Proceedings of the 34th European conference on Advances in Information Retrieval. Barcelona，Spain. 2012：522-526.

卫冰洁(1987—)，博士，中级工程师，主要研究领域为微博检索及数据挖掘。
E-mail：weibingjie1986@163.com

王斌(1972—)，博士，研究员，主要研究领域为信息检索及自然语言处理。
E-mail：wangbin@iie.ac.cn

张帅(1987—)，硕士，工程师，主要研究领域为微博分类及信息检索。
E-mail：zhangshuai01@ict.ac.cn

(上接第 9 页)

李业刚(1975—)，博士研究生，副教授，主要研究领域为自然语言处理，机器翻译。
E-mail：lyg8256@bit.edu.cn；lyg8256@qq.com

黄河燕(1963—)，博士，教授，主要研究领域为自然语言处理与机器翻译。
E-mail：hhy63@bit.edu.cn

史树敏(1978—)，博士，讲师，主要研究领域为自然语言处理，本体方法论及应用。
E-mail：bjssm@bit.edu.cn