情报工程 ISSN 2095-915X
第2卷 第4期 80-88 2016年8月

TECHNOLOGY INTELLIGENCE ENGINEERING
ISSN 2095-915X Vol.2 No.4 80-88 Aug 2016

doi:10.3772/j.issn.2095-915x.2016.04.010

统计机器翻译领域自适应方法比较研究

丁亮，李颖，何彦青

（中国科学技术信息研究所 北京 100038）

摘要：统计机器翻译常常面临训练数据与待翻译文本领域不一的问题，从而影响了翻译的性能，因此领域自适应一直是研究者关注的课题。本文以传统自适应方法和现行的机器学习方法为框架，介绍了近年来统计机器翻译领域自适应研究的进展。分析了各类研究方法的优缺点并对未来研究做出展望。

关键词：统计机器翻译，领域自适应，语料选取，翻译性能改进

中图分类号：G35，TP391.41

Comparison Study of Domain Adaption Methods in Statistical Machine Translation

DING Liang，LI Ying，HE YanQing

(Institute of Scientific and Technical Information of China, Beijing 100038, China)

Abstract: Statistical machine translation（SMT）is often faced with the problem of different domains between the training data set and test data set,which affecting the performance of translation,therefore domain adaptation has been a subject of concern.In this paper,we constructed the domain adaptation research framework-the traditional adaptive methods and the existing machine learning methods,and introduced the research progress in recent years.We analyzed the advantages and disadvantages of each method and made a prospect for the future research.

Keywords: Statistical machine translation,domain adaptation,corpus selection,translation performance improvement

基金项目：本文受国家自然科学基金项目：（61303152、 71503240 和 71403257），中国科学技术信息研究所重点工作项目：（ZD2016-05）资助。

作者简介：丁亮（1994-），硕士研究生，主要研究方向：机器翻译，自然语言处理；李颖（1964-），博士，副研究员，主要研究方向：数字图书馆技术、知识工程、语言技术；何彦青（1974-），博士，副研究员，主要研究方向：机器翻译，自然语言处理，Email：heyq@istic.ac.cn。

1 引言

国家主席习近平于 2013 年提出共建“丝绸之路利用计算机程序实现源语言到目标语言的翻译过程被称为机器翻译，在信息时代人们对于不同语言之间的翻译需求也在明显增长，机器翻译是自然语言理解的一个重要应用。机器翻译的发展经历过曲折的过程，上世纪五十年代经历了繁荣期，由于人们对机器翻译有不切实际的期望而受到客观条件限制，翻译质量一直没有提高。1966 年美国发布 ALPAC 报告打击了人们对于机器翻译的热情，之后机器翻译进入了冬天。七十年代，随着计算语言学和人工智能的发展，机器翻译的研究工作开始复苏。八九十年代随着新的有效翻译方法的提出，机器翻译进入了第二个繁荣期。

机器翻译主要分为传统的基于规则的方法和基于语料的机器翻译方法，其中基于语料的机器翻译方法又分为统计机器翻译方法和基于实例的翻译方法。统计机器翻译（Statistical Machine Translation）系统通常在句子对齐的双语对译语料（下文简称训练数据）上训练，利用该语料的词对齐学习翻译规则，通过对数线性模型进行寻优生成目标翻译。该机制中影响翻译质量的因素有很多，其中包括训练数据的领域分布、句子对齐质量和句对规模。一般来说，训练数据与测试数据的领域越接近、句子对齐质量越高、句对数量越多，越有助于从中学习到更加精准的翻译规则，从而获取更为鲁棒的译文。在实际应用中，为了追求训练数据的质量和规模，训练数据通常会来源繁杂，主题多样，文体不一，与待翻译的目标文本的领域并不能保证完全一致，因而产生了“领域自适应”问题。在很多统计机器翻译系统中，领域自适应问题成为了最主要的待解决问题之一。

统计机器翻译的领域自适应，其目标在于筛选或者规划训练数据，以及设计和调整翻译模型，使得统计机器翻译系统能为待翻译的文本生成更符合其领域特性的翻译结果。在信息爆炸的时代，机器翻译面对的翻译任务越来越复杂，挑战中就包括数据来源于不同的领域的问题，比如新闻、军事、财经、生物等等，同时在相同领域内也面临着文体不一的现象，这样的差异导致了统计机器翻译中领域间自适应和领域内自适应两种问题。领域间表现在采用某一领域训练数据训练的翻译模型针对该领域的测试集会有优异的翻译表现，但对于其他领域的待翻译文本，翻译效果往往不佳；领域内自适应问题表现在，如果我们采用泛化性较强的多领域训练数据训练翻译模型，那么针对特定的领域翻译表现就会欠佳。诸多学者通过传统的机器学习方法已经提出一些文本分类器模型，包括基于语义分析的分类模型、半监督的分类模型以、集成文本分类模型以及实时分类模型，但是大部分都是简单的分类或者回归，对于统计机器翻译尝试方法较少，这也是当前统计机器翻译领域的研究热点。

2 统计机器翻译背景

对于统计机器翻译，根据 Brow 的定义 [1]，给定源语言句子 f，由统计机器翻译系统计算最大化翻译概率，得到最佳译文 e_{best}。

$$\begin{aligned} e_{best} &= \operatorname{argmax}_e p(e/f) \\ &= \operatorname{argmax}_e p(e/f) p_{\mathrm{LM}}(e) \end{aligned} \qquad \text{（公式 1）}$$

统计机器翻译系统通过大规模的双语平行语料来训练翻译模型 $p(f/e)$，并且通过目标语言端语料训练语言模型 $P_{Lm}(e)$。Och 等人 [2] 引入基于最大熵模型的统计机器翻译方法，最大熵模型增强了翻译系统的泛化性，可以同时考虑多个特征。假设分别是 e、f 上的 M 个特征，λ_1，…，λ_{M} 分别是每个特征对应的权重，数学描述为：

$$e_{best} = \arg\max_{e} \frac{\exp \sum_{m=1}^{M} \lambda_m h_m(e,f)}{\sum_{e'} \exp \sum_{m=1}^{M} \lambda_m h_m(e',f)} \quad （公式 2）$$
$$= \arg\max_{e} \{\sum_{m=1}^{M} \lambda_m h_m(e,f)\}$$

在 Och 提出的最大熵方法的 SMT 中，翻译过程主要有特征选择和参数训练，为翻译过程中选取的特征分配不同的权重。

一个标准的基于短语的统计机器翻译系统通常包括训练、调优和翻译三个阶段的处理，见图 1，需要准备训练数据、单语目标语言语料、开发集和测试集。训练数据为双语对译语料，多为句子对齐，经过预处理和词对齐后，获取各种翻译规则，包括短语翻译表、调序概率以及最大熵调序参数等。单语的目标语言的语料，可以使用训练数据的目标语言端，也可以再额外添加更多的单语数据，多为句子级别，用来训练语言模型。除了训练过程中生成的各种翻译规则和语言模型之外，解码器的运行还需要特征权重，调优的过程就是在开发集上对特征权重进行选择。开发集是源语言的句子集合，每个源语言句子带有一个或多个目标语言的参考翻译。在开发集上调优，通常使用最小错误训练，需要解码器不断迭代当前特征参数，通过自动计算并比较 BLEU 打分，再改变权重重新解码，直到达到迭代次数上限或者翻译系统表现稳定为止，这是一个多维参数优化的问题。利用训练过程的翻译规则、语言模型和调优得到的特征权重，解码器就可以实现翻译过程，使用测试集来进行翻译并进行 BLEU 打分，从而观察翻译系统的翻译效果。

3 领域自适应研究

当前最常用的统计机器翻译系统大多都采用大规模的双语语料和基于短语的翻译模型，在统计机器翻译系统上训练出翻译模型和 n-gram 语言模型，采用最小错误训练进行解码。对于训练的翻译模型，若测试数据和该翻译模型领域信息比较一致，则可获得较好的译文，反之，若测试数据与训练语料领域不一，则译文质量就会有所下降，且出现非行业术语的翻译情况时，会造成很大的理解不便。事实上，要为所有测试集找到领域一致的平行语料并不容易，通过领域自适应来解决这一问题是相对可行的。统计机器翻译经常面临训练数据与待翻译文本的领域不一致的问题，进而影响翻译的性能，因此领域自适应一直是研究者关注的课题。

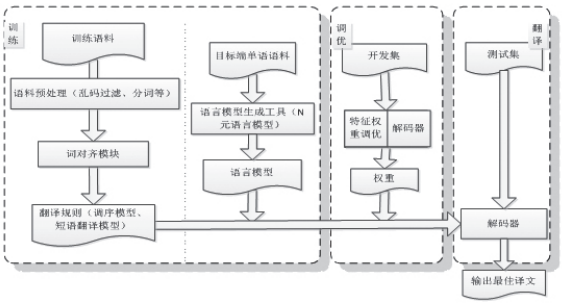

图 1 翻译系统流程图

已有的统计机器翻译领域自适应方法包括：一种是传统方法，包括数据自适应和混合模型方法。二是现行的机器学习方法，包括半监督学习、主题模型。在传统的自适应研究方法中，数据自适应的方法通过设计相似度函数，在训练数据中选择和待翻译测试集文本领域更为一致的源领域数据，这种领域更为“相似”的数据选择方法不仅可以缩小训练规模，也能有效的提高翻译质量。子模型自适应的方法将训练数据根据不同的领域特征划分为几类，针对原训练集的不同子集分别训练子翻译模型，然后为每个子翻译模型调整权重，使翻译系统针对不同领域的待翻译文本仍然有较好的翻译表现。近几年随着机器学习方法的兴起，也有学者将半监督学习和主题模型的研究思路引入领域自适应，半监督学习方法将每次新的测试句子及其翻译结果再次组成双语句对，放回训练数据重新训练翻译系统，增大训练

语料中测试集领域句子的比重，每一次迭代都是给翻译模型正确样本的过程，这种半监督的方法一直迭代到翻译性能稳定为止，有效提升了译文质量达到领域自适应目的。基于主题模型（Topic model）的方法则将文本的主题信息融合在统计机器翻译的训练或者解码的过程中，用于提升翻译性能。

3.1 基于传统方法的领域自适应

传统方法主要是通过对数据和模型的研究、选取达到领域自适应的目的。基于数据自适应方法利用相似度函数选择与目标领域文本相似的源领域数据来进行模型训练。

Eck 等人 [3] 采用 N-gram 的频次和信息检索中的 TF-IDF 来选择语言模型数据，使用新的测试集的译文重新选择语言模型的训练数据，在西班牙 – 英语的训练语料上取得了性能改善，但是改进效果并不明显且只能在给定的语料和测试集上进行实验，很难推广。

Zhao 等人 [4] 同样采用信息检索中 TF-IDF 方法，将新测试集的译文的语言模型与训练集的语言模型进行插值，通过对语言模型的自适应以达到领域自适应目的，在四个类别的汉英语料上进行训练，并取得了译文质量的改善，BLEU 打分有了明显的提升，该方法难免脱离难推广的缺点。

吕雅娟等 [5] 提出离线翻译的方法，通过 TF-IDF 为训练数据中的每一个双语句对用重复累加的方式赋以权重，与测试集更接近的句子赋予更高的权重，这种方法扩大了训练语料的规模，同时达到了数据领域自适应的目的，该实验通过离线的方式进行不同领域的模型训练，在线上完成测试文本的类别识别与模型选择，虽然翻译性能得到改善，达到了领域自适应，但是增加了翻译成本以及系统复杂度。

Matsoukas 等 [6] 则使用判别式模型对训练数据赋权重，用阿拉伯语 – 英语语料进行实验，通过对多类别的训练语料通过参数估计为每个句子加上权值，实现了翻译模型的领域自适应，该方案取得了译文质量的提升，但也大幅度增加了额外的训练时间。另外，交叉熵也可以用来选择语言模型数据或者双语训练数据，通过交叉熵来区分句子的类别，但在实验中只简单的区分了有领域词和无领域词两种，因此制作出了简单的二值分类问题，虽然取得了试验效果的提升，并没有借鉴已有的语义网络和标签进行领域区分。

姚树杰等人 [7] 利用词语或者短语的覆盖率，并设计了一套对数线性模型来进行语料的过滤，这种方法也可以用来选择训练数据，姚的论文只采用了部分训练数据就达到了与原来语料规模可比较的翻译结果，但也没有加上明确的语义标签进行类别区分。

之前的方法大多利用测试集来与训练语料进行相似度计算并筛选语料，Zheng 等人认为 [8]，针对既定的训练语料，相对于设计复杂的算法进行数据筛选，模型参数的训练也非常重要，他们的实验选择与测试集领域相似的开发集进行参数调优，同样有很好的翻译性能改善，这些方法大都根据测试集，来改善训练样本，以达到领域自适应目的，不同之处在于相似度函数的设计以及所处理的数据集。

基于子模型自适应的方法更适合在线的机器翻译，此类方法将训练数据根据不同的领域特征分为几个不同的子集，利用每个子集的数据分别训练翻译子模型，再根据测试数据的上下文信息适当地为每个子模型调整权重。

文献 [9] 按照训练数据的不同来源将数据进行分类，利用测试数据与每一个子训练数据的文本距离分配子翻译模型的权重，不但计算了 TF-IDF 和困惑度（Perplexity），还利用了浅层语义分析（Latent Semantic Analysis）和 EM 技术。

吕雅娟等 [5] 提出的在线翻译也是混合模型，通过对不同短语表中的短语概率进行插值以选择

最适合测试数据的具体模型，在翻译表现上得到了改善，但并没有给出确定的语义标签。

Jorge Civera 和 Alfons Juan 于 2007 年 [10] 提出一种改变数据密度的混合模型，将隐马尔科夫对齐模型进行了改进，为普通的 HMM 对齐模型引入 T 参数用以混合不同模型，并将 $p(t|y)$ 作为混合模型的权值，如下：

$$p(x|y)=\sum_{t=1}^{T}p(t|y)p(x|y,t)$$
$$=\sum_{t=1}^{T}p(t|y)\sum_{a\in A(x,y)}p(x,a|y,t) \quad \text{（公式 3）}$$

在西班牙 – 英语欧洲新闻语料上进行了混合模型的实验，实验结果均有一定的提升，但是在加大了训练复杂度的情况下，只取得了微弱的改进 , 且该方法并没有给出语料类别明确的标签。

Koehn 等 [11] 利用最小错误率训练的方法调整混合模型的参数，该论文最大的贡献在于提出最小错误训练，现在主流的统计机器翻译系统 moses、Niutrans 等都采用最小错误训练，在时间复杂度和译文质量都有很大的改善。

Finch 和 sumita 等 [12] 针对不同类型的句子，如疑问句和陈述句训练混合模型，在训练样本上按照不同的句式对语料进行分类训练以适应测试集，这种方法在一些社区问答系统上比较实用，但是其局限性也导致了推广和应用的难度。

Foster 等 [13] 使用 logistic 模型对短语表的特征加权 ，通过每个句子在翻译模型中概率表上的权值变化达到领域自适应的目的，在极大程度上改善了翻译质量，增加了系统复杂性且无明确的语义标签来区分类别。

BaneIjee 等 [14] 利用混合模型翻译在线论坛的内容，他们采用线性对数加权的方式进行领域之间的区分，并设计了多组对照试验，用基准系统和进行过领域自适应的系统进行对比，翻译质量有了一些改善，但其应用比较局限。

Sennrich 等 [15] 使用最小化困惑度（perplexity minimization）调整混合模型的参数，该研究对于翻译模型、语言模型都进行了探讨，关于采用几个特征、特征参数的求解都进行了研究，最后实验部分展示除了性能的改进，但是并未给出明确的语义标签对类别进行区分，很难在之后借鉴或者形成固有的知识库。

Hal Daume III 等人的研究 [16] 认为翻译系统移植到新的领域后导致的翻译错误主要归咎于未登录词（Out Of Vocabulary，OOV）、集外词，因而主张从目标领域的可比语料中挖掘词典来解决 OOV 术语的翻译问题，她们设计了挖掘未登录词的实验对翻译系统进行了改进，并整合进了统计机器翻译系统，以德法语料作为样本，将领域信息分为新闻、欧洲药品代理文本、open subtitles 字幕信息和 PHP 技术文档四类，通过手工选取每个领域高频未登录词并挖掘同语义词类构建未登录词词典来扩充训练数据，使得实验结果整体得到提升，其中欧洲药品代理文本特征最强，效果也最好，这种方法虽然使得翻译性能得到提升，但是起初的四类文本都是手选，该方法并不具有普适性，难以推广。

Wuebker J 等人 [17] 在 2015 年提出一种层级的增量自适应方法，该方法将权重和规则根据每次增加的语料进行更新，实验效果表明该方法对于改善翻译质量有明显的提升，但是也很难用该方法构建语义标签知识库进行类别区分。这些基于混合模型的方法都没有针对数据的主题内容进行训练数据的切分，重点大多放在翻译子模型的参数调整上。

3.2 基于机器学习方法的领域自适应

一些学者尝试用统计机器学习的方法来解决机器翻译领域自适应的问题，比较有效的有半监督或无监督自学习方法和主题模型方法。通过自学习来增强翻译系统对于特定领域的翻译表现，达到领域自适应的目的。

半监督自学习方法借鉴了机器学习中此类方法的机制，通过不断地将源语言的单语文本经机器翻译后得到的高质量翻译再放回训练数据，不断迭代重新训练翻译系统。

Ueffing 等人[18]采用直推式半监督学习（Transductive learning）方式更加有效的利用了目标语言源语言端，提出几种算法对照不同的细节对于翻译质量的改善，实验采用了法语－英语以及中英两大类实验数据，均得到改善，但是半监督的方法都难以改进的是其无具体的语义标签，因此这种方法很难被之后的工作借鉴。

百度的吴华等[19]使用领域外数据训练翻译系统，再用目标领域翻译词典和单语语料来改善领域内的翻译表现，并提出一种融合算法框架如图2所示，这种翻译方法曾应用于百度翻译的翻译技术上，并取得了比较好的翻译表现。

Schwenk 等人[20]用初始翻译系统翻译大规模单语数据来改善翻译系统，此类方法虽然得到了性能提升，但是没有给领域数据进行具体的标记。

Lambert P 等人[21]提出无监督训练方式，使用单语数据实现自我增强从而适应翻译模型，通过这种自我增强的自适应学习可以达到语言模型平滑的效果，但是仍然无法给不同类别的语料加上标签。此类半监督或者无监督方法实现了统计机器翻译的领域自适应并改善了翻译质量，但没有对领域标签进行具体的定义。

Input 领域外训练数据L_O
领域内翻译词典D_I
领域内目标语言语料T_I（可选项）
领域内源语言语料S_I（可选项）
Begin 为D_I分配翻译概率
If T_I 可获取
Training step: $\pi = Estimate(L_O, D_I, T_I)$, *where* π represents the general model
Else
Training step: $\pi = Estimate(L_O, D_I)$
End if
If S_I 可获取
$\pi^{(0)} = \pi$， i = 0
Re *peat*
$i = i + 1$
Labeling step: *Translate* S_I *with* $\pi^{(i-1)}$ *to get synthetic bilingual corpus* L_I
Training step: $\pi^{(i)} = \mathrm{Re}-estimate(L_O, D_I, L_I)$
Until 不能得到改进
$\pi = \pi^{(i)}$
End if
End
Output Model π 用于领域内翻译

图2 领域自适应算法

主题模型（topic models）使用词（word）和文档（document）的共现矩阵来对文档集（text）建立生成模型来推断主题。使用主题模型可以将文档以一定的概率聚类到给定的主题上，通过主题自动获取词间关系。

Zhao B 等人[22]首次使用隐马尔可夫模型和双语主题混合模型改善了词对齐的准确性，Zhao B 等人的研究对基于主题的词汇翻译模型进行了估计，提升了机器翻译的性能。

Tam 等人[23]认为在训练中加强一对一主题对应，可以有效地将一种语言的浅层主题分布转移到另一种语言上，通过对统计机器翻译中的语言模型和翻译词典同时进行自适应来提升机器翻译效果。这些研究都是在词级翻译上使用了主题信息。

Su 等[24]利用目标领域单语文本的主题信息，对基于短语的翻译模型进行了领域自适应。Xiao 等[25]则通过构建层次短语翻译规则的主题信息模型，在解码过程中创建主题相似度，进行层次短语规则的选取。这两个研究均是在短语级别上使用了主体信息。主题模型考虑了文本的主题信息，该主题多为文档集中自动训练获取，特别是无监督学习算法，虽然都实现了领域自适应并改进了译文质量，但没有主题信息的显性表达。

4 研究展望

近年来领域自适应研究方法在统计机器翻译领域已经取得很多进展，翻译质量都可以通过这些研究得到提升，但大都是以测试数据为基准，着重于对训练数据或者翻译模型进行领域的适应调整，都没有给出训练数据或者测试数据明确的领域标签，当更换了测试数据之后，则需要重新进行领域自适应。如果能够给出测试数据或者训练数据明确的领域标签，利用领域标签对各种数据分门别类地处理，再分别训练各领域的翻译模型，那么，即使更换了测试数据，也只需要对测试数据进行领域归类，再根据领域类别选择翻译模型进行翻译，这样更适合维护统计机器翻译系统，有利于数据积累和长期规划。一方面，作者建议可以借鉴 WordNet 或者 HowNet 等其他的语义知识，或者双语领域映射关系来完善领域标注，这样既能避免机器学习方法中半监督和无监督自适应方法中的领域标签问题。另一方面，未来研究中可以借鉴深度学习中 ffnnlm（Feed-forward Neural Net LanguageModel），通过训练词向量神经网络语言模型来表征语义信息，从而尝试领域自适应的进一步研究。

参考文献

[1]Brown P F, Pietra V J D, Pietra S A D, et al. The Mathematics of Statistical Machine Translation: Parameter estimation[J]. Computational Linguistics, 1993, 19(2):263-311.

[2]Och F J, Ney H. Discriminative training and Maximum Entropy Models for Statistical Machine Translation[C]// Meeting on Association for Computational Linguistics. Association for Computational Linguistics, 2002:295–302.

[3] Eck M, Vogel S, Waibel A. Low Cost Portability for Statistical Machine Translation Based on N-gram Coverage[J]. Proceedings of Mtsummit X, 2005:1-7.

[4]Zhao B，EcK M，VOGEL S. Language Model Adaptation for Statistical Machine Translation via Structured Query Models[C],Proceedings of Coling 2004. Geneva，Switzerland:COLING，2004：411-417.

[5]L ü Y, Huang J, Liu Q. Improving Statistical Machine Translation Performance by Training Data Selection and Optimization[C]// EMNLP-CoNLL 2007, Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, June 28-30, 2007, Prague, Czech Republic, 2007:343-350.

[6]Matsoukas S, Rosti A V I, Zhang B. Discriminative Corpus Weight Estimation for Machine Translation[C] // Proceedings of the 2009 Conference on Empirical Methods in Natrual Language Processing. Singapore:Association for Computational Linguistics, 2009:708-717.

[7] 姚树杰，肖桐，朱靖波．基于句对质量和覆盖度的统计机器翻译训练语料选取 [J]. 中文信息学报，2011, 25(2):72-77.

[8]Zheng Z, He Z, Meng Y, et al. Domain Adaptation for Statistical Machine Translation in Development Corpus Selection[C]// Universal Communication Symposium （IUCS）, 2010 4th International. IEEE, 2010:2-7.

[9] Foster G， Kuhn R. Mixture model adaptation for SMT[C] // Proceedings of Second Workshop on Statistical Machine Translation. Prague，Czech Republic：Association for Computational Linguistics，2007:128-135.

[10]Civera J，Juan A. Domain Adaptationin Statistical Machine Translation with Mixture modelling[C] // Proceedings of the Second workshop Statistical Machine Translation. Prague，Czech Repllblic： Association for Computational Linguistics，2007:177-180.

[11]Koehn P，Schroeder J. Experiments in Domain Adaptation for Statistical Machine Translation[C] // Proceedings of the second. Workshop on Statistical Machine Translation . Prague，Czech Republic：Association for Computational Linguistics，2007:224-227.

[12]Finch A, Sumita E. Dynamic model Interpolation for Statistical Machine Translation[C]//Proceedings of the Third Workshop on Statistical Machine translation. Columbus，Ohio：Associationfor Computational Linguistics，2008：208-215.

[13] Foster G,，Goutte C，Kuhn R. Discriminative Instance Weighting for Domain Adaptation in Staitistical Machine Translation[C]//Proceedings of the 2010 Conference on Empirical methods in natural language processing, Cambrige， MA: Association for Computational Linguistics，2010：451-459.

[14]Banerjee P,Naskar S,Roturier J,et al. Domain Adaptation in Statistical Machine Translation of User-forum Data using Component-Level Mixture Modeling[J]. Proceedings of the 13th Machine Translation Summit, 2011: 285-292.

[15]Sennrich R. Perplexity Minimization for Translation Model Domain Daptation in Statistical Machine Translation[C]//Proceedings of the 13th Conference of the EuropeanChapter of the Association for Computational Linguistics . Avignon，France:Association for Computational Linguistics，2012:539-549.

[16] Hal Daumé III, Jagadeesh Jagarlamudi, Domain Adaptation for Machine Translation by Mining Unseen Words[C]//Proceedings of the 49th ACL:shortpapers,Portland, Oregon: Association for Computational Linguistics，2011:407-412.

[17]Wuebker J, Green S, Denero J. Hierarchical Incremental Adaptation for Statistical Machine Translation[C]// Conference on Empirical Methods in Natural Language Processing, 2015:1059-1065.

[18] Ueffing N, Haffari G, Sarkar A. Semi-supervised Model Adaptation for Statistical Machine Translation[J]. Machine translation, 2007, 21:71-94.

[19]Wu H，Wang H，Zong C. Domain Adaptation for Statistical Machine Translation with Domain Dictionary and Monolingual corploa[C] // Pmceedings of the 22nd Intenational conference on computational Linguistics （Coling 2008）. Manchester，UK:Coling 2008 Organizing Committee, 2008:993-1000.

[20]Schwenk H. Investigations on Large-scale Lightly Supervised Training for Statistical Machine trallslation[C]//Proceedings of theInternational Workshop on Spoken Language Translation. Hawaii, USA:IWSLT,2008:182-189.

[21]Lambert P, Schwenk H, Servan C, et al. Investigations on Translation Model Adaptation Using Monolingual Data[C]// Empirical Methods in Natural Language Processing / Workshop on Statistical Machine Translation. 2011:284-293.

[22]Zhao B, Xing E P. BiTAM:Bilingual Topic

Admixture Models for Word Alignment[C]// Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions. Sydney, Australia: Association for Computational Linguistics, 2006: 969-976.

[23]Tam YC, Lane I, Schultz T. Bilingual LSA-based Adaptation for Statistical Machine Translation[J]. Machine Translation, 2007, 2l(4): 187-207.

[24]Su J, Wu H, Wang H, et al. Translation Model Adaptation for Statistical Machine Translation With Monolingual Topic Information[C]//Proceedings of Annual Meeting of the Association for Computational Linguistics:Human Language Technologies.Jeju, Korea:Association for Computational Linguistics, 2012:459-468.

[25]Xiao X, Xiong D, Zhang M, et al. A Topic Similarity Model for Hieralchical Phrase-based Translation[C]//Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Jeju, Korea: Association for computational Linguistics, 2012:750-758.