doi:10.3772/j.issn.2095-915x.2017.03.005

基于枢轴语言的平行语料构建方法

北京交通大学计算机与信息技术学院 北京 100044
单华 张玉洁 周雯 徐金安 陈钰枫

摘 要 平行语料库的规模对于统计机器翻译性能的提高具有重要作用，但是平行语料库的人工构建成本很高。针对这个问题，本文提出了一种低成本高效率的平行语料构建方法，利用枢轴语言作为桥梁，借助已有的机器翻译技术并融合主动学习方法构建目标语言对的大规模高质量平行语料库。本文通过以英语作为枢轴语言构建日汉平行语料库的实例研究，利用成熟的基于短语的统计机器翻译技术，描述了基于译文自动评测的良好译文选择方法、基于主动学习的语料选取方法、以及翻译系统的更新迭代和评价实验。实验结果表明，本文提出的方法能够快速构建日汉平行语料，并有效提高日汉翻译系统的性能。

关键词： 枢轴语言，机器翻译，平行语料，主动学习
中图分类号： G35

Approach of Constructing Parallel Corpus Based on Pivot Language

The School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044, China
SHAN Hua ZHANG YuJie ZHOU Wen XU JinAn CHEN YuFeng

Abstract A large scale parallel corpus plays an important role in improving the performance of machine translation. It spent highly for manually constructing a parallel corpus. This paper proposed a pivot based approach for constructing high quality parallel corpus with low cost, in which the existing machine translation technology and active learning method are combined. This paper describes the domain adaptation method based on active learning, the good translation selection method based on automatic translation evaluation, and iterative retraining of translation system. We applied the approach to the construction of Japanese-

基金项目： 本文受国家自然科学基金(61370130, 61473294)的资助。
作者简介： 单华（1990–），硕士，研究方向：自然语言处理，机器翻译，Email:13120422@bjtu.edu.cn；张玉洁（1961–），博士，教授，研究方向：自然语言处理，机器翻译，文本大数据处理；周雯（1993–），硕士，研究方向：自然语言处理，信息抽取；徐金安（1970–），博士，副教授，研究方向：自然语言处理，机器翻译；陈钰枫（1981–），博士，副教授，研究方向：自然语言处理，机器翻译。

doi:10.3772/j.issn.2095-915x.2017.03.005

Chinese parallel corpus by taking English as pivot and conducted evaluation experiments. The experimental results showed that the proposed approach effectively obtained Japanese-Chinese parallel corpus with high quality and the constructed parallel corpus indeed improved the performance of Japanese-Chinese machine translation system.

Keywords：Pivot language, machine translation, parallel corpus, active learning

1 引言

在机器翻译领域，英语与其他语言的机器翻译一直得到大量的研发投入，积累了大规模的平行语料等翻译资源，译文质量也达到了一定实用水平。但是其他语言对之间的机器翻译得到的关注相对较少，特别是在科技领域等具有大量需求的特定领域中，平行语料等翻译资源匮乏，机器翻译质量尚未达到实用水平。因此，在开发面向特定领域机器翻译的任务中，现有的问题一方面是如何将平行语料库从小规模扩大到大规模，另一方面是如何将机器翻译系统从已有领域扩展到特定领域。

传统的获取平行语料的方法有人工翻译和利用机器翻译译文的两种方法，前者成本高效率低，而且需要高质量的翻译人员；后者效率较高，省时省力，但受翻译系统性能影响，平行语料质量低。于是，研究人员开始考虑机器翻译与人工校正相结合的方法来构建平行语料，并成为切实可行的方法。其中，如何减少人工校正同时获得高质量的平行语料成为关注的焦点[1]。

针对这些问题，本文提出了一种低成本高效率的平行语料构建方法，利用英语作为中间语言，借助已有的机器翻译技术构建目标翻译语言对的平行语料，达到提高机器翻译系统性能的目标。本课题涉及三个基本问题：1）使用机器翻译构建的平行语料对改善翻译系统的性能是否有帮助；2）自动译文的质量会有噪音，如何过滤噪音获取良好译文；3）如果需要构建特定领域的平行语料，如何利用较少人工校正提高翻译系统的领域适应能力。针对这些问题，我们提出利用译文自动评测方法对译文进行评分，选取质量良好的译文；并基于主动学习（Active Learning）方法提出利用语言模型选取少量域外语句的策略，以解决机器翻译系统的领域适应问题。

本文通过以英语作为枢轴语言构建日汉平行语料库的实例开展研究。由于日英、汉英机器翻译的研发历史较长，我们可以获得大规模高质量的日英和汉英平行语料。另一方面，日汉双语人才很少，直接构建日汉语料的成本会很高，需要借助英语这一枢轴语言，因为英汉双语人才较多，可以为这一方法的实施提供保障。

本文第 2 节介绍了利用机器翻译技术构建平行语料的相关研究；第 3 节描述了基于枢轴语言的平行语料构建框架、基于译文自动评测的良好译文过滤方法和基于主动学习的领域适应方法；第 4 节通过实验对所提方法的性能进行验证，并与已有方法进行比较；第 5 节给出本文的结论，并提出今后研究方向。

2 相关研究

针对机器翻译中平行语料匮乏的问题，研究人员尝试利用枢轴语言构建平行语料库。Li 等在日汉专利平行语料构建中，利用英语作为枢轴语言，借助 Google 英汉翻译引擎将已有的日英平行语料中的英文句子翻译成汉语，获得日汉平行语料 [2]。但是该方法在从语料中选取语句翻译时只考虑了语句之间在单词粒度上的差异性，并未参考实际译文的质量以及上下文信息。

Utiyama 等采用了首先翻译日语语句得到 n-best 英语译文，再将其翻译成汉语的作法 [3]。但是这种方法需经过两次机器翻译，会导致翻译错误传播与扩大，使平行语料含有更大噪音。

Wu 则利用多个翻译系统翻译枢轴语言，在得到的多个系统译文中选取质量最好的译文，在对某一个系统的译文进行评价时，采用了将其他系统译文作为参考译文的作法 [4]。但是该方法并没有考虑欲构建的平行语料与已有平行语料在语言现象上的差异性。

需要构建平行语料的另一种情形是已有机器翻译系统的领域适应问题。某个领域的训练语料上搭建的机器翻译系统，在面向不同领域时，通常翻译质量会有所下降。人们开始关注语料之间在语言现象上的差异性，并提出利用主动学习方法发现已有机器翻译系统在面对新领域时的不足之处，并针对性地增加翻译知识。具体地，从新领域大规模语料库中选取蕴含最多新领域信息量的句子，对其翻译并进行人工校正得到新领域的语句对；然后将其加入训练语料重新训练翻译系统 [5-8]，在开发跨领域机器翻译系统时显示了较好性能。因此，如何选择蕴含信息量多的句子成为该方法的关键。Eck 等通过抽取含有未登录 n-gram 的句子进行人工翻译，并将其加入平行语料，以提高训练集的覆盖率 [5]。其缺陷是选取语句的方法无法选取低概率的 n-gram。Bloodgood 等提出从源语言单语料中选出相对于已有平行语料的未登录 n-gram 并按概率排序，选择高位的 n-gram 进行人工翻译 [6]。但是，由于没有考虑已有翻译系统实际的译文质量，可能使人工翻译成为无效的工作。

利用机器翻译构建平行语料的另一个问题是机器译文中的噪音，即质量不好的译文。Ananthakrishnan 等提出错误驱动的方法，通过在开发集上选择翻译错误最多的句子发现已有翻译系统的不足 [7]。Haffari 等利用未登录 n-gram 的数量、与已有平行语料的相似性以及译文置信度作为特征训练分类器，判断一个新句对中的源语言句子能否提供最大信息量，并且其译文质量是否达标 [8]。

我们发现，前人在利用枢轴语言构建平行语料的工作中通常选用现成的翻译系统，当领域不同时很难保证译文质量，而他们只是通过选取相对较好的译文缓解这一问题，导致获得的平行语料很难覆盖新领域的语言现象。因此，我们认为在基于枢轴语言的平行语料构建方法中，中间翻译系统的性能至关重要。所以，本论文提出了融合主动学习的基于枢轴语言的平行语料构建方法，通过主动学习扩展枢轴语言 - 目标语言在特定领域上的平行语料，以提高中间翻译系统的译文质量；并提出基于译文自动评测的良好译文选择方法。

doi:10.3772/j.issn.2095-915x.2017.03.005

3 融合主动学习的基于枢轴语言的平行语料构建方法

本节具体介绍融合主动学习的基于枢轴语言的平行语料构建方法，以英语作为枢轴语言构建日汉平行语料库为实例，对引言中提到的三个基本问题开展研究，描述具体解决方案，试图满足在领域变换时将平行语料库从小规模扩大到大规模、以及将已有机器翻译系统扩展到特定领域的需求。

3.1 框架概述

为了构建特定领域上的日汉平行语料，我们将目光转向易于获取的相同领域的大规模日英平行语料，借助英汉翻译系统获得汉语译文。那么，为了能够高质量地翻译特定领域上的英语句子，我们采用主动学习方法实现英汉翻译系统的领域适应。总体框架如图 1 所示。我们有领域 A 上的小规模日汉平行语料搭建的日汉翻译系统，为了翻译新领域 B 上的文本，需构建领域 B 上的日汉平行语料，这是我们的目标，在图中下部用长横线标识。同时，我们拥有领域 B 上的大规模日英平行语料以及领域 C 上的大规模英汉平行语料。为此，我们借助领域 C 上开发的英汉翻译系统翻译领域 B 上日英平行语料的英语部分，得到汉语译文，从而构成领域 B 上的日汉平行语料，在图中中部用短横线标识。为了得到较好的汉语译文，我们利用主动学习方法将英汉翻译系统从领域 C 适应到领域 B，在图中上部用点 - 横线标识。其中，基于主动学习的句子选取模块从领域 B 上的日英平行语料选取相对于领域 C 上的英汉平行语料差异性最多的英语句子优先进行译后校正；基于译文自动评测的译文选取模块对汉语译文进行评分排序，选择质量好的译文与其对应的日语句子构成平行句对。

在利用主动学习将英汉机器翻译系统进行领域适应的过程中，首先进行基于主动学习的英文句子选取，将得到的英文句子借助英汉机器翻译系统翻译成汉语，并对汉语译文进行人工校正，得到英汉平行句对；然后并入领域 C 上的英汉平行语料重新训练英汉翻译系统，并用测试集评价。重复上述操作直到测试集上的性能没有明显提高时停止。当然，也可根据投入的人工成本设置迭代次数。

在利用英语为枢轴语言构建日汉平行语料时，需要借助英汉翻译系统翻译目标领域的日英平行语料。此时要考虑英汉翻译系统的领域和目标领域的差异性，若两个领域一致，则可以取得良好的译文；若不一致，则需要利用主动学习方法解决英汉翻译系统的领域适应问题。我们可以借助 n-gram 语言模型计算两个领域的训练语料之间的差异性，以判断领域一致的程度。

3.2 基于译文自动评测的良好译文选取方法

在利用枢轴语言构建平行语料的方法中，不可避免的问题是自动译文中存在翻译错误，直接加入平行语料会导致训练集含有大量噪音。为此，我们设计基于译文自动评测的汉语译文选取方法，过滤噪音，提高平行语料的质量。在评测过程中，我们考虑了忠实度（adequacy）和流利度（fluency）两个指标。

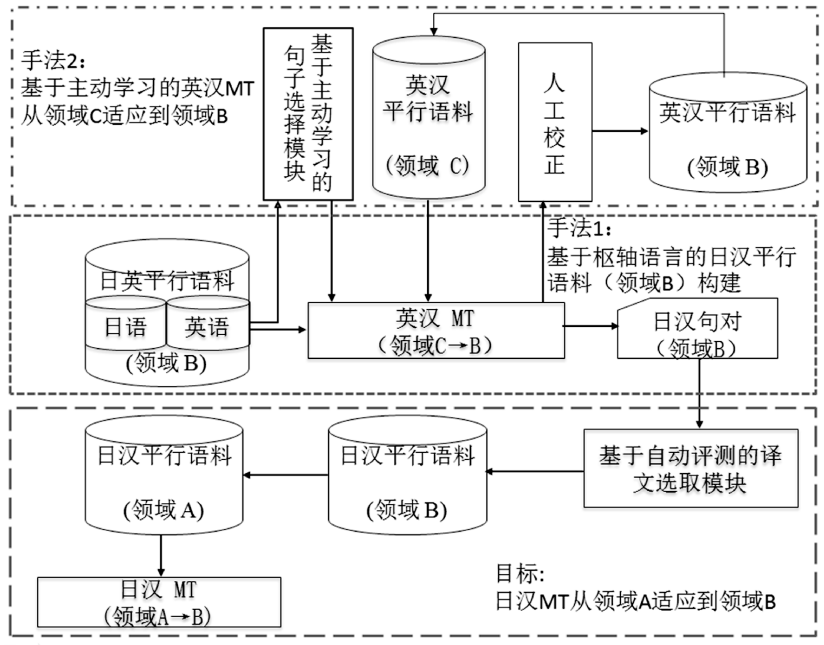

图1　融合主动学习的基于枢轴语言的平行语料构建方法的整体框架

对于忠实度，在无参考译文的情况下，我们借鉴了二次翻译的方法[9]，得到其原文句子的回译。具体地，这里待评测的译文是英汉翻译系统输出的汉语译文，预先使用同样的训练语料开发一个汉英翻译系统，利用该汉英翻译系统对一次翻译中的汉语译文进行二次翻译，得到英语译文，并通过英文译文对英文原句的忠实度的评价来评测汉语译文的忠实度。我们通过将英语原文作为参考译文，分别采用TER[10]与OrthoBleu[9]进行评测。其中，TER表示将译文修改成参考译文所需的最小修改率，即翻译错误率（Translation Error Rate），由公式（1）定义。OrthoBleu是一种基于字符的编辑距离方法，主要解决TER不能处理的复合词和缩写词造成的评测误差，由公式（2）定义。

$$\text{TER}=\frac{\#\text{单词编辑操作数}}{\#\text{参考句子单词数}} \quad （公式1）$$

$$\text{OrthoBleu}=\frac{\#\text{字符编辑操作数}}{\#\text{参考句子字符数}} \quad （公式2）$$

对于流利度，本文使用了一种基于N-gram的无参考译文的方法。根据汉语的语言模型，以句子长度作为惩罚，对每个汉语译文计算评分，得分较高的译文我们认为它的质量也高。在后面的实验部分，我们将会介绍三种评分策略的具体实现和对它们的评价。

3.3　基于主动学习的领域适应方法

在本文提出的方法中，为了获得高质量特定领域上的汉语译文，本文采用主动学习的方法，低成本高效率地提高英汉翻译系统的领域适应能力。为此，本文根据新领域的日英平行语料与训练英汉翻译系统的英汉平行语料之间在英文句子上的语言现象的差异性进行选取。选取方法主要基于以下考虑：一方面，英语的n-gram在日英语料中出现的频率越高，需要翻译的量越大，因此在训练语料中增加这样的平行句对越重要；另一方面，英语的n-gam在英

doi:10.3772/j.issn.2095-915x.2017.03.005

汉语料中出现的频率越高，翻译的质量会越好，因此在训练语料中再增加这样的平行句对的意义或价值会降低。具体地，我们采用了基于语言模型的句子选取策略[8]选取英文句子，如公式（3）所示。

$$\phi^N(s) := \sum_{n=1}^{N} \frac{w_n}{|X_s^n|} \sum_{x \in X_s^n} \log \frac{P(x|U,n)}{P(x|L,n)} \quad \text{（公式 3）}$$

$f^N(s)$ 为句子 的 n-gram 的加权和，即被选取的评分。X_s^n 表示句子 s 的 n-gram 集合，$P(x|D,n)$ 是在语料 D 上 n-gram 集合中的概率，权重 w_n 调整不同长度的 n-gram 分数的重要性。在主动学习方法的描述中 L 表示带有标注信息的语料，U 表示未标注信息的语料，通过计算 U 中 s 的 n-gram 与 L 之间的差异性选择 s 进行标注。在我们的任务中，L 对应已有训练语料英汉平行语料的英语部分，U 对应新领域的日英平行语料的英语部分，标注信息就是新领域的日英平行语料对应的汉语译文。那么，主动学习的目的就是从 U 中，即新领域中选取与已有训练语料 L 差异性大的英文句子优先标注其汉语译文，具体地，对译文进行人工校正。评分越高说明该句子与已有语料的差异性越大，越需要优先选取。本文使用的是 N=4 的语言模型，对应的权重向量 w^T 设置为（0.15,0.2,0.3,0.35）[8]。

首先计算排序并选取高位的固定数量的英文句子构成集合 U^+；利用当前 时的英汉翻译系统 $MT^{(t)}$ 进行翻译并进行人工校正，获得英汉平行语料 U_c^+；用 $U_c^+ \cup L$ 作为新的训练集重新训练得到更新后的英汉翻译系统 $MT^{(t+1)}$；此时，$L=U_c^+ \cup L$，$U=U-U_c^+$。如此迭代地重复上述操作，直到翻译系统的性能没有明显提高。

4 实验设计及结果分析

为了验证本文提出的平行语料构建方法，我们通过专利文献的特定领域上基于英语构建日汉平行语料的实验，对本文方法在第 1 节提到的三个基本问题上的应用效果分别进行验证。

4.1 实验设计及数据

在 3.1 节，我们定义了领域 A 上的日汉平行语料，领域 B 上的日英平行语料以及领域 C 上的英汉平行语料。在本文实验中我们分别对应地选取 ASPEC①日汉平行语料、NTCIR-10②日英平行语料和 NTCIR-10 英汉平行语料（这三个语料的来源不同，可以认为来自不同领域，我们没有对它们的领域进行判定），为方便起见，用 A、B、C 代表它们各自的领域。为了验证本文方法在平行语料匮乏的实际情景中的效果，本文从三个语料中随机抽取小规模数据进行模拟实验，具体使用的数据规模如表 1 所示。

在日汉平行语料上我们搭建了领域 A 上的日汉翻译系统，为了评测该系统在从领域 A 切换到领域 B 时在性能上的差异，我们分别利用这两个领域的测试集对系统进行评测。领域 A 上的测试集使用 ASPEC 的日汉测试集，命名为 Test_JC；关于领域 B 上的测试集，我们通过对 NTCIR-10 日英平行语料中测试集进行人工翻译得到对应的汉语译文，命名该测试集为 Test_EJC。此数据也将用于评测英汉翻译系统在领

①http://orchid.kuee.kyoto-u.ac.jp/ASPEC/
②http://research.nii.ac.jp/ntcir/ntcir-10/

域 B 上的适应效果。为了使日汉翻译系统能够适应领域 B，我们需利用英汉翻译系统翻译 NTCIR-10 日英平行语料构建领域 B 上的日汉平行语料。为此，我们利用 NTCIR-10 的英汉平行语料搭建了领域 C 上的英汉翻译系统，并利用主动学习方法使该英汉翻译系统适应领域 B，随后利用该系统翻译 NTCIR-10 日英平行语料获得领域 B 上的日汉平行语料。

本论文统一采用 Moses③搭建统计机器翻译系统，用 BLEU[12] 值对系统性能进行评测。

表1　实验数据说明

语料	训练集数目（句对）	开发集数目（句对）	测试集数目（句对）
英 - 汉（NTCIR-10，领域 C）	17000	2000	2000
日 - 英（NTCIR-10，领域 B）	50000	2000	2000
日 - 汉（ASPEC，领域 A）	12000	2090	2107

4.2　基于枢轴语言构建平行语料方法的实现及评测

首先，我们分析搭建好的日汉翻译系统（Baseline）在领域切换后的性能变化。为此，我们用同一领域的 Test_JC 测试集和新领域的 Test_EJC 测试集分别对系统进行评测并比较，实验结果如表 2 所示。值得注意的是，与其他研究的评测结果相比，本文的 BLEU 值偏低，原因是我们只使用了极少量的训练语料。由于随后 4.4 实验中的人工校正译文数量不可能太多，我们相应地缩小了训练语料的数量，这也同时缩短了系统搭建时间。实验结果显示，Baseline 系统在领域切换后，BLEU 值下降 73%。

表2　不同领域间日汉翻译系统的性能比较

测试集	Test_JC（领域 A）	Test_EJC（领域 B）
BLEU	0.1985	0.0540

然后，我们开始验证第一个问题，即使用机器翻译构建的平行语料对改善翻译系统的性能是否有帮助。为此，我们利用搭建的英汉翻译系统翻译领域 B 上的日英平行语料的英语部分，得到汉语译文。这里将获得的领域 B 上的 50,000 句对的日汉平行语料称为粗糙日汉平行语料。随后将其加入领域 A 的 ASPEC 日汉平行语料，重新训练得到日汉翻译系统 Raw，并在 Test_EJC 上进行评测，评测结果如表 3 所示。结果显示 BLEU 值由 0.054 增加到 0.146，提高了 73%。由此，我们认为以英语作为枢轴语言，借助机器翻译构建新领域的平行语料的方法确实能够有效改善在领域切换中大幅下降的翻译系统的性能。

表3　使用基于枢轴语言构建的平行语料前后日汉翻译系统的性能比较

系统	Baseline	Raw
BLEU	0.0540	0.1460

4.3　基于译文自动评测的语句选取方法的实现及评测

上面实验中得到的粗糙日汉平行语料含有被错误翻译的汉语译文。接下来，我们使用 3.2 节提出的基于译文自动评测的语句选取方法过滤汉语译文，选取高位的 25,000 句汉语译文及其对应的日语构成良好平行语料；随后将其加入领域 A 的 ASPEC 日汉平行语料，重新训练翻译系统，验证该方法的有效性。同样，利用

③http://www.statmt.org/moses/

doi:10.3772/j.issn.2095-915x.2017.03.005

低位的 25,000 句也进行了实验。下面描述根据三种不同评分策略过滤数据后搭建的三个系统。

Ngram_sort：使用基于语言模型的策略进行过滤，具体地，使用 SRILM④在英汉平行语料的汉语部分进行训练，得到汉语语言模型，并用该语言模型对汉语译文进行评分和排序 [11]。

TER_sort：使用 TER 进行过滤，即计算二次翻译得到的英语译文与英语原文之间的 TER 并排序。这里，为了二次翻译，我们在 NTCIR-10 汉英平行语料上搭建了汉英翻译系统。

OrthoBleu_sort：使用 OrthoBLEU 进行过滤，即计算二次翻译得到的英语译文与其英语原文之间的 OrthoBLEU 并排序。

然后，我们在 Test_EJC 测试集上对上述三个系统进行评测，并与 Raw、Baseline 系统的性能进行比较，评测结果显示在后面的表 4 中。其中，括号中的数字是利用低位的 25,000 句构成的系统的评测结果。对表 4 中的结果分析如下。

在基于自动译文评测方法搭建的三个系统中，使用 OrthoBLEU 策略搭建的系统的 BLEU 值最高，另外两个系统的 BLEU 值相同，略低于 OrthoBleu_sort。显示了 OrthoBLEU 策略在汉语译文选取中具有一定优势。

Ngram_sort、TER_sort、OrthoBleu_sort 系统与 Raw 系统相比，BLEU 值十分接近，仅相差 2%，但是只用了一半的数据作为训练语料。该结果表明，利用译文的自动评测方法的确能够选取质量较好的平行语料，并且可以用来替换全部的粗糙平行语料，充分显示了我们所提出的基于译文自动评测方法在构建平行语料中的有效性。

Ngram_sort、TER_sort、OrthoBleu_sort 系统的 BLEU 值与括号中的结果相比，基于高位译文的翻译系统都获得了更好的性能，说明三个评分策略在良好译文选取上的有效性。

4.4 基于主动学习的领域适应方法的实现及评测

在以英语作为枢轴语言构建日汉平行语料的任务中，英汉翻译系统的质量至关重要。上述实验显示，使用基于译文自动评测的方法可以有效过滤汉语译文中的噪音。接下来，我们实现 3.3 节描述的基于主动学习的英汉翻译系统的领域适应方法，并对该方法进行验证。具体地，每次从英日平行语料选取 250 句英文进行翻译和校正、以及翻译系统的重新训练，共迭代 6 次。所以，一共人工校正了 1,500 句英语句子的汉语译文。然后，利用最后一次迭代更新后的英汉翻译系统将剩余的 48,500 句英语翻译成汉语，共得到 50,000 句对的日汉平行语料，将这些语料并入 ASPEC 日汉平行语料重新训练日汉翻译系统，记为 AL。

首先，我们对英汉翻译系统领域适应的效果进行验证。为此，通过在 Test_EJC 上的评测结果观察迭代过程中 BLEU 值的提升幅度，评测结果如图 2 所示。从图中可以看出 BLEU 值随着每次迭代都有明显提升，最终得到了 12% 的提高，说明我们所提的方法对于改善英汉翻译系统的性能是有效的。

随后，我们在 Test_EJC 测试集上评测 AL

④http://www.statmt.org/moses/

系统，得到的 BLEU 值为 0.1621。对比前面的实验结果，日汉翻译系统的 BLEU 值得到很大提升，相比于 Raw 提高了 11%。对比结果表明经过领域适应后的英汉翻译系统的确提高了所构建的日汉平行语料的质量，验证了基于主动学习的领域适应方法的有效性。

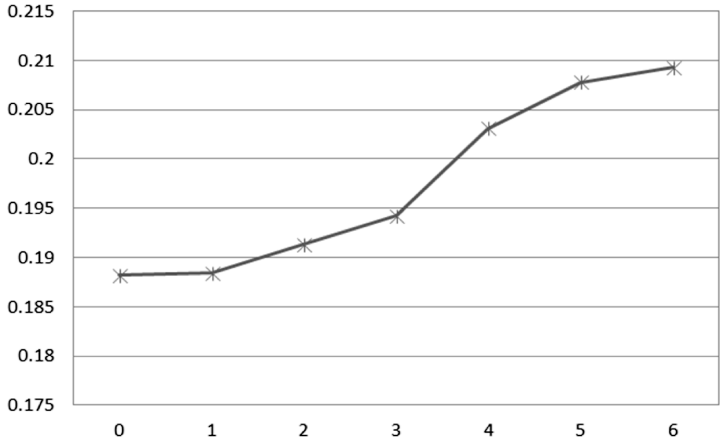

图2　在基于主动学习的迭代过程中英汉翻译系统的BLEU值

表4　不同方法构建的平行语料上所搭建的日汉翻译系统的性能比较

系统	Baseline	Raw	基于自动译文评测方法搭建的系统			AL
			Ngram_sort	TER_sort	OrthoBleu_sort	
BLEU	0.0540	0.1460	0.1424（0.1365）	0.1424（0.1368）	0.1438（0.1381）	0.1621

上面的实验结果显示，日汉翻译系统在加入新构建的平行语料之后，BLEU 值得到了明显提高。接下来，我们从语言模型覆盖率的角度分析性能提高的原因 [13]，并预估系统 BLEU 值提升的空间。为此，我们考察日汉翻译系统的 ASPEC 平行语料（领域 A）对于同一领域上的测试集 Test_JC 以及新领域 B 上 NTCIR-10 的测试集 Test_EJC 的覆盖率，分析它们之间的差异性；然后考察加入新构建的领域 B 的平行语料之后对 Test_EJC 的覆盖率，分析新构建的平行语料对覆盖率提升的贡献度。具体地，我们针对各个平行语料中的日语部分进行计算，计算结果显示在表 5 中。

表5　从语言模型评测日汉训练语料变化前后对测试集的覆盖率

训练语料（相对于测试集）	1-gram/%	2-gram/%	3-gram/%	4-gram/%	5-gram/%
ASPEC 覆盖同领域 A 的测试集 Test_JC	83.72	55.03	52.22	38.96	27.22
ASPEC 覆盖不同领域 B 的测试集 Test_EJC	76.04	35.66	21.09	11.54	5.79
ASPEC+ 新构建领域 B 语料覆盖 Test_EJC	93.02	75.52	64.91	47.12	35.66

从表 5 中可以发现，加入新构建的日汉平行语料后，对测试集 Test_EJC 的覆盖率得到了明显提升，而且比原有 ASPEC 日汉平行语料对其同一领域上测试集 Test_JC 的覆盖率还要高。

doi:10.3772/j.issn.2095-915x.2017.03.005

前面的实验结果显示，翻译系统 Baseline 翻译相同领域上的测试集的 BLEU 值为 0.1985。由此，我们期待加入新构建的平行语料后翻译系统的 BLEU 值应该达到或超过这个水平。但是，目前的 BLEU 值才达到 0.1621。基于这个分析结果，我们推断本次实验中人工校正的数据还不足够多，英汉翻译系统的领域适应还不充分，导致汉语译文的错误给平行语料带入噪音，影响了系统性能的提升。由此看出在基于枢轴语言构成高质量平行语料的方面还有提升的空间。

5 结语

本文提出了融合主动学习的基于枢轴语言的平行语料构建方法和基于自动评测的译文选取方法，并应用于以英语作为枢轴语言构建日汉平行语料的任务中。评测结果显示，1）基于自动译文评测方法所获得的 BLEU 值的提升接近直接使用机器译文的方法，而使用语料的数量是后者的一半，表明本文所提出的译文选取方法的有效性；2）基于主动学习的领域适应方法在英汉翻译系统上 BLEU 值提高了 12%，利用其所构建的日汉平行语料使日汉翻译系统在 BLEU 值上提高了 11%，说明了本文所提出的领域适应方法的有效性。本文所描述的方法可以用于低成本高效率地构建其它语言对的平行语料。

本论文对基于主动学习的领域适应方法和基于自动评测的译文选取方法分别进行了单独的实验，今后，我们将考虑结合两个方法构建平行语料。在基于自动评测的译文选取方法中，本论文的方法是选取一定数量的排在高位的平行语料进行实验，今后我们将进行详细实验寻找其他的选择标准。

参考文献

[1] Zaidan O F, Callison-Burch C. Crowdsourcing Translation: Professional Quality from Non-Professionals[C]// The Meeting of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference, 19-24 June, 2011, Portland, Oregon, USA. DBLP, 2011: 1220-1229.

[2] Li X, Meng Y, Yu H. Improving Chinese-to-Japanese Patent Translation using English as Pivot Language[J]. Faculty of Computer Science Universitas Indonesia, 2012.

[3] Utiyama M, Isahara H. A Comparison of Pivot Methods for Phrase-Based Statistical Machine Translation[C]// Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, April 22-27, 2007, Rochester, New York, USA. DBLP, 2007: 484-491.

[4] Wu H, Wang H. Revisiting Pivot Language Approach for Machine Translation[C]// ACL 2009, Proceedings of the, Meeting of the Association for Computational Linguistics and the, International Joint Conference on Natural Language Processing of the Afnlp, 2-7 August 2009, Singapore. DBLP, 2009:154-162.

[5] Eck M, Vogel S, Waibel A. Low Cost Portability for Statistical Machine Translation based in N-gram Frequency and TF-IDF[J]. In proceedings of International Workshop on Spoken Language Translation (IWSLT), 2005.

[6] Bloodgood M, Callison-Burch C. Bucking the Trend: Large-Scale Cost-Focused Active Learning for Statistical Machine Translation[C]// ACL 2010,

Proceedings of the, Meeting of the Association for Computational Linguistics, July 11-16, 2010, Uppsala, Sweden. DBLP, 2010: 854-864.

[7] Ananthakrishnan S, Prasad R, Stallard D, et al. Discriminative Sample Selection for Statistical Machine Translation[C]// Conference on Empirical Methods in Natural Language Processing, EMNLP 2010, 9-11 October 2010, Mit Stata Center, Massachusetts, USA, A Meeting of Sigdat, A Special Interest Group of the ACL. DBLP, 2010: 626-635.

[8] Haffari G, Roy M, Sarkar A. Active Learning for Statistical Phrase-based Machine Translation[C]// Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, May 31 - June 5, 2009, Boulder, Colorado, USA. DBLP, 2009: 415-423.

[9] Rapp R. The Back-translation Score: Automatic MT Evaluation at the Sentence Level without Reference Translations[C]// ACL 2009, Proceedings of the, Meeting of the Association for Computational Linguistics and the, International Joint Conference on Natural Language Processing of the AFNLP, 2-7 August 2009, Singapore, Short Papers. DBLP, 2009: 133-136.

[10] Snover M, Dorr B, Schwartz R, et al. A study of Translation Edit Rate with Targeted Human Annotation[J]. Machine Translation Workshop North Bethesda Md, 2006(1): 223-231.

[11] Gamon M, Aue A, Smets M. Sentence-Level MT Evaluation without Reference Translations: beyond Language Modeling[C]// European Association for Machine Translation. 2005.

[12] Papineni K, Roukos S, Ward T, et al. BLEU: A Method for Automatic Evaluation of Machine Translation[C]// Meeting on Association for Computational Linguistics. Association for Computational Linguistics, 2002: 311-318.

[13] Doddington G. Automatic Evaluation of Machine Translation Quality using n-Gram Concurrence Statistics [EB/OL]. [2017-01-13]. http://www.itl.nist.gov/iad/mig/tests/mt/doc/ngram-study.pdf.