基于二维词汇化领域知识的日汉科技术语翻译方法研究*

丁亮[1] 李颖[1] 何彦青[1**] 刘建辉[2]

（1.中国科学技术信息研究所，北京 100038；2.河北地质大学,石家庄，050031）

摘要：科技术语翻译要求高度的准确性和专业性，通过建立术语语料的领域知识标签，并基于待翻译术语的领域对训练语料进行筛选，可训练出针对领域的翻译模型，能极大改善科技术语翻译质量。这对于机器翻译、词典自动编纂、跨语言信息检索等自然语言处理任务都具有重要的实用价值。本项研究以日汉科技术语翻译为研究目标，通过自动获取日语术语的二维领域知识，包括语义范畴标签和应用场景标签，可以对日语术语进行领域归类。该方法应用到统计机器翻译中，可以自动标注日语术语（或句子）的二维领域知识，生成测试集和开发集的领域标签集合，进而筛选训练数据。实验表明，仅利用部分训练数据就可以获取与原始训练数据可比较的翻译结果，验证了本方法的有效性和可行性。本研究有助于解决面向科技术语的统计机器翻译的领域自适应问题。

关键词：统计机器翻译；训练语料选取；领域自适应

Research on Japanese-Chinese S&T Terminology Translation Based-on Two-dimensional Domain Lexicalized Domain Knowledge

Ding Liang[1]，Li Ying[1]，He Yanqing[1]，Liu Jianhui[2]

（ 1.Institute of scientific and Technical Information of China, Beijing 100038； 2. Hebei GEO University，Shijiazhuang,050031）

Abstract: Scientific and Technical terminology translation requires higher accuracy and professionalism. Through the establishment of the domain knowledge labels of the terminology, it can select the training corpus and train the translation model with adaptive domains, which can greatly improve the translation quality of scientific and technical terminology has important practical value for machine translation, compilation of bilingual dictionaries, cross-language information retrieval, etc. The study aims at Japanese-Chinese terminology translation by automatically accessing two-dimensional lexicalized domain knowledge of Japanese term, such as semantic category and application scenarios. Based on those information Japanese terminology can be classified into different domains. When applied to statistical machine translation, our

---

基金项目：本研究受国家自然科学基金项目（No.61303152、No.71503240 和 No.71403257）和中国科学技术信息研究所重点工作项目（No.ZD2016-05）资助。

作者简介：丁亮（1994—），男，硕士研究生，主要研究方向：机器翻译，自然语言处理；李颖（1964—），女，中国科学技术信息研究所副研究员，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—），女，中国科学技术信息研究所副研究员，主要研究方向：机器翻译，自然语言处理；刘建辉（1969—），男，河北地质大学副教授，主要研究方向：语言学、文学理论。

** 通讯作者，地址：海淀区复兴路 15 号 邮箱：heyq@istic.ac.cn

approach annotate the domain labels for Japanese terms (or sentences) and then generate domain label set of test-set and development-set to filter training data. Our experimental results show that only a part of the training data can gain a comparative translation performance with the whole training data. This verifies that the approach is efficient and feasible and helps to solve the domain adaption problem of statistical machine translation.

Keywords: machine translation ； training data selection ； domain adaption

中图分类号：TP391　　**文献标识码：**A

1 引言

术语是各专业领域中特定概念的语言表示，是借助语言形式表达或限定专业概念的约定性符号[1]。术语的翻译是将一种语言表达中的术语转化为另一种语言的对应词汇。科技文献中蕴含了大量的术语，翻译要求高度的准确性和专业性。因此，科技术语翻译的自动获取研究既有很大的技术难度又有重要的实用价值。

从既有的研究成果来看，术语的自动翻译有两种生成方式：一种术语翻译的方式是抽取对齐法，这种方法直接从数据源中执行抽取过程来识别双语术语，然后进行对齐，即对两种语言中已经识别的术语建立源语言术语到目标语言术语之间的链接关系。抽取对齐方法可以保证目标术语符合人类的语言习惯，其难点则在于术语的边界确定以及对应关系的保证。另一种是直接翻译方法，这种方法与命名实体（Named Entity）的自动翻译相类似，是利用词典或者平行语料建立机器翻译模型的方式来生成源语言的目标翻译。目前最流行的模式是统计机器翻译（Statistical Machine Translation）[2]。由于科技术语的专业属性很强，所覆盖的领域又很宽泛，而统计机器翻译非常依赖训练数据，因此经常会产生“领域自适应”问题。这个问题的解决需要筛选或者规划训练数据，或者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术语生成更符合其领域特性的翻译结果。

已有的统计机器翻译领域自适应方法包括：基于数据选择的方法、基于混合模型的方法、半监督学习方法和基于话题模型的方法[3]。基于数据选择的方法通过设计相似度函数，选择和目标领域文本“相似”的源领域数据；基于混合模型的方法将训练数据划分为几部分，分别训练子翻译模型，然后为每个子翻译模型调整权重；半监督学习方法将测试句子及其翻译结果组成双语句对，放回训练数据重新训练翻译系统，一直迭代到翻译性能稳定为止；基于话题（Topic）模型的方法则将话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。这些方法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适应。

本研究目标是面向日汉科技术语的统计机器翻译。为了充分利用科技术语的专业属性，借助语义范畴标签和应用场景标签，从纵横两个维度来形成二维领域知识，进而获取术语(或者句子)级别的领域知识。借助该领域知识，采用基于数据选择的方法进行语料过滤，即分别对训练集、开发集和待翻译的测试集获取术语(或者句子)级别的二维领域知识，合并得到目标领域知识标签集合来筛选训练数据，从而保证领域的一致性。本研究在统计机器翻译系统上进行了测试，仅利用部分训练数据就获取了与原始训练数据可比较的翻译效果。该研究对于机器翻译、词典自动编纂、跨语言信息检索等自然语言处理任务都有重要的实用价值。

2 相关工作

术语翻译有两类获取方法，其一为抽取对齐方法，其二为直接翻译方法。前者直接从网络或者语料中找到互为对应翻译的术语对抽取出来，后者通过建立规则或者翻译模型直接对源语言的术语进行翻译。抽取对齐方法的来源可以是网络数据[4]，也可以是可比较语料[5]或平行语料[6]。

由于篇幅有限，这里不多赘述。直接翻译方法可以通过人工进行，也可以采用词典或者机器翻译来获取。人工翻译成本太高，而词典又无法保证大量的未登录词汇的准确处理。因此人们更多地采用机器翻译方法,特别是统计机器翻译。统计机器翻译来翻译术语的时候，需要解决领域自适应问题。

领域自适应方法可以分为基于数据选择的方法、基于混合模型的方法、自学习为代表的半监督学习方法和基于话题模型的方法四大类[3]。基于数据选择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训练模型训练。某些文献[7,8]采用信息检索中的 TF-IDF 来选择语言模型数据，吕雅娟等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赋以权重[9]。这些方法不同之处在于相似度函数的设计以及所处理的数据集。基于混合模型的方法适合在线机器翻译，此类方法将训练数据分为几个不同的部分，利用每一部分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。Koehn 等利用最小错误率训练的方法调整混合模型的参数[10]。Finch 和 Sumita 针对不同类型的句子，如疑问句和陈述句训练混合模型[11]。这些基于混合模型的方法都没有针对数据的话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。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，然后再放回训练数据，不断迭代重新训练翻译系统。Ueffing 采用直推式半监督学习（Transductive learning）方式[12]。吴华等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[13]。话题模型（topic models）使用词和文档（document）的共现矩阵来对文档集（text）建立生成模型来推断话题。使用话题模型可以将文档以一定的概率聚类到给定的话题上，通过话题自动获取词间关系。有些文献[14, 15]使用隐马尔可夫模型和双语话题混合模型改善了词对齐的准确性，并提升了机器翻译的性能。Xiao 等则通过构建层次短语翻译规则的话题信息模型，在解码过程中创建话题相似度，进行层次短语规则的选取[16]。与上述三种方法相比，话题模型考虑了文本的主题信息，但是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，并没有主题信息的显性表达。

本研究方法首次把日语二维领域知识引入到统计机器翻译中，对日语词汇分别标注语义范畴标签和应用场景标签，整合词汇级别的领域知识形成术语级别的领域知识，对日语术语的领域做了显性的表达，通过测试数据集与开发集的领域知识对原始训练语料进行筛选，在保证翻译质量的同时有效减少了训练数据规模与训练成本，并达到了日汉科技术语翻译中训练数据与测试数据领域自适应的目的。

3 基于领域知识的统计机器翻译训练数据选择

下面分别介绍日语科技术语的二维领域知识标注方法、基于短语的统计机器翻译系统，以及利用术语的二维领域知识过滤训练数据的算法。

3.1 日语二维领域知识标注

科技术语通常文本较短、专业性强，多为名词。日文的名词词汇可以按照语义范畴和应用场景两个维度进行分类。语义范畴有 22 种，见表 1，其中有的词可以一词多个语义范畴标签，如“鱼”，“蔬菜”可以分别归类为“动物”、“植物”，也可以表示“人工物—食物”。应用场景设定了 12 种，见表 2。每个日语词可能有多个应用场景标签，比如：“大学院（大学院）”可以属于＜教育学习＞类，也可以为 ＜科学・技术＞类，“円の切り上げ（日元升值）”既属于＜经济＞类又属于 ＜政治＞类。

由于大多数名词词汇都有其语义范畴标签与应用场景标签，故本研究将这两种分类合起来视为二维领域知识标签。

在对日语句子进行二维领域知识标注时，首先对句子进行分词和词性标注，然后利用领域词典对每个日语基本词汇标注

表 1： 语义范畴标签集及示例

日文（中文翻译）	例子	日文（中文翻译）	例子
人(人物)	学生,老师,歌手	自然物（自然物体)	石头,砂子,地下水
組織・団体(组织团体)	政府,军,国家	場所-施設（场所—建筑)	门,天安门
動物(动物类)	狗,怪兽,宠物	場所-施設部位（场所—建筑部分)	窗户,围墙,走廊
植物(植物)	树木,樱花,绿藻	場所-自然（场所—自然)	山,海洋,天空
動物-部位(动物—部分)	手,皮肤,伤	場所-機能（场所—功能)	上,边缘,境界
植物-部位(植物—部分)	叶,孢子,落叶	場所-その他(场所—其他)	市,村,战场
人工物-食べ物(人造物品—食物)	料理,便当,饲料	抽象物(抽象物体)	理由,能力,语言
人工物-衣類(人造物品—衣物)	毛衣,衬衫,饰品	形・模様(形状花纹)	圆形,正方形,大型
人工物-乗り物(人造物品—交通工具)	汽车,轮椅,火箭	色(颜色)	红色,青色,黄色
人工物-金銭(人造物品—金钱)	工资,保险金,压岁钱	数量(数量)	复数,多数,和
人工物-その他(人造物品—其他)	铅笔,杯子,眼镜	時間(时间)	年,月,早晨

语义范畴和应用场景，这里的基本词汇多为名词和动词。图 1[17]为半自动构建领域词典的过程。即先自动构建一个领域词典表示每个日语基本词汇与领域关键词的语义关系，然后人工修正该词典。自动构建分 4 步：

1）为每个领域手工选定一些关键词，该过程对应图 1 中第一步，为设定好的每个 $DOMAIN_i$ 手工选定大约 20 个关键词 $\mathrm{kw}_{ik}(k=a,b,c...)$；

2）为了构建词典，选取网络中的高频词作为基本词，基本词汇的集合即图 1 中第二步的 $JFW_s\{JFW_1,JFW_2,JFW_3...\}$，为每个基本词汇与领域 $DOMAIN_i$ 之间计算一个关联得分 A_d， A_d 为每个领域中得分最高的前五个 A_k 之和；基本词和领域的关键词之间的关联得分为 A_k，目的为计算第 j

表 2 应用场景标签集及示例

日文	中文	例子
文化・芸術	文化艺术	照片,电影,音乐
レクリエーション	娱乐	游乐场,游戏,游玩
スポーツス	体育	选手,比赛,运动
健康・医学	健康医学	医疗,医院,患者
家庭・暮らし	家庭生活	搬家,住宅,家庭
料理・食事	饮食	零食,吃饭,料理
交通	交通	车站,道路,运行
教育・学習	教育学习	教授,报告,入学
科学・技術	科学技术	学会,分析,理论
ビジネス	经济（商业）	贩卖,商品,经营
メディア	媒体	播放,报纸,节目
政治	政治	行政,政府,军队

个基本词 JFW_j 与第 i 个领域中第 k 个关键词 kw_{ik} 的关联程度，此处采用 χ^2 卡方统计量来计算 A_k 并使用待标记日文科技术语（或句子）作为语料。记 n 是日语待标记句子总行数，若待标记的当前日文术语（句子）中基本词和关键词的共现次数，则记为 $con(JFW \&\& kw)$，若只包含基本词或者关键词，则出现次数分别记为 $con(JFW)$ 或 $con(kw)$，基本词（JFW）和关键词（kw）之间的关联得分 A_k 计算方式如下：

$$A_k(JFW,kw)=\frac{n(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)}$$

其中

$$a=con(JFW \&\& kw),\quad b=con(JFW)-a$$
$$c=con(kw)-a,\quad d=n-(a+b+c).$$

3）计算基本词和每个领域的得分 A_d，并选取得分最高的领域作为该基本词汇 JFW 的相关领域,为每个基本词都迭代这个过程，得到结果如图 1 第三步；

4）扩展基本词汇与多领域的关系。为每个基本词设置领域阈值，将满足阈值的领域标签都赋予该基本词汇，如果大于阈值的领域不止一个，这样会出现一词多领域；如果领域关联得分都小于阈值，则出现没有领域（$NODOMAIN$）的情况，其中语义范畴标签的词典构建同应用场景标签词典构建类似，不再赘述。

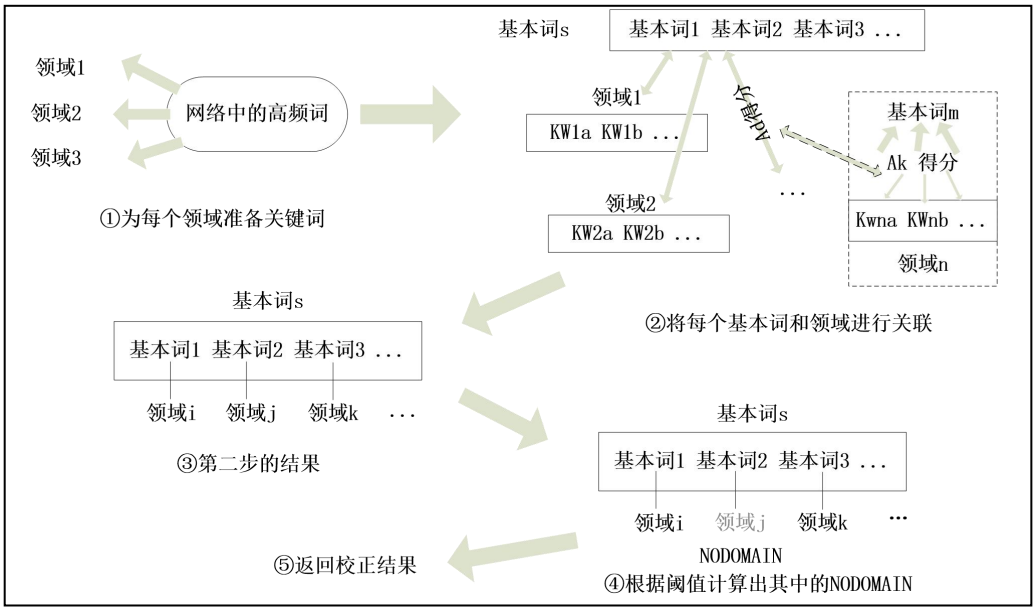

图 1 半自动构建领域词典的过程[17]

3.2 基于短语的统计机器翻译模型

在基于短语的统计机器翻译[18-20]中，给定源语言句子 $S=s_1s_2...s_L$,通常采用对数线性模型来实现翻译，其过程就是搜索具有最大概率的目标翻译 $T=t_1t_2...t_K$：

$$T^*=\arg\max_T\sum_{m=1}^{M}\lambda_m h_m(S,T)$$

其中 $h_m(S,T)$ 为特征函数，λ_m 为特征权重。这里基于短语的翻译系统使用了以下特征函数：

1) 基于相对频率的短语后验概率

2)词汇化短语后验概率

3)语言模型特征

4)词惩罚

5)短语惩罚

6) 基于最大熵的词汇化重排序特征

7) MSD 重排序特征

3.3 利用二维领域知识筛选训练语料

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 2，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，这里使用测试集来进行翻译并进行 BLEU 打分，从而评价翻译系统的翻译效果。

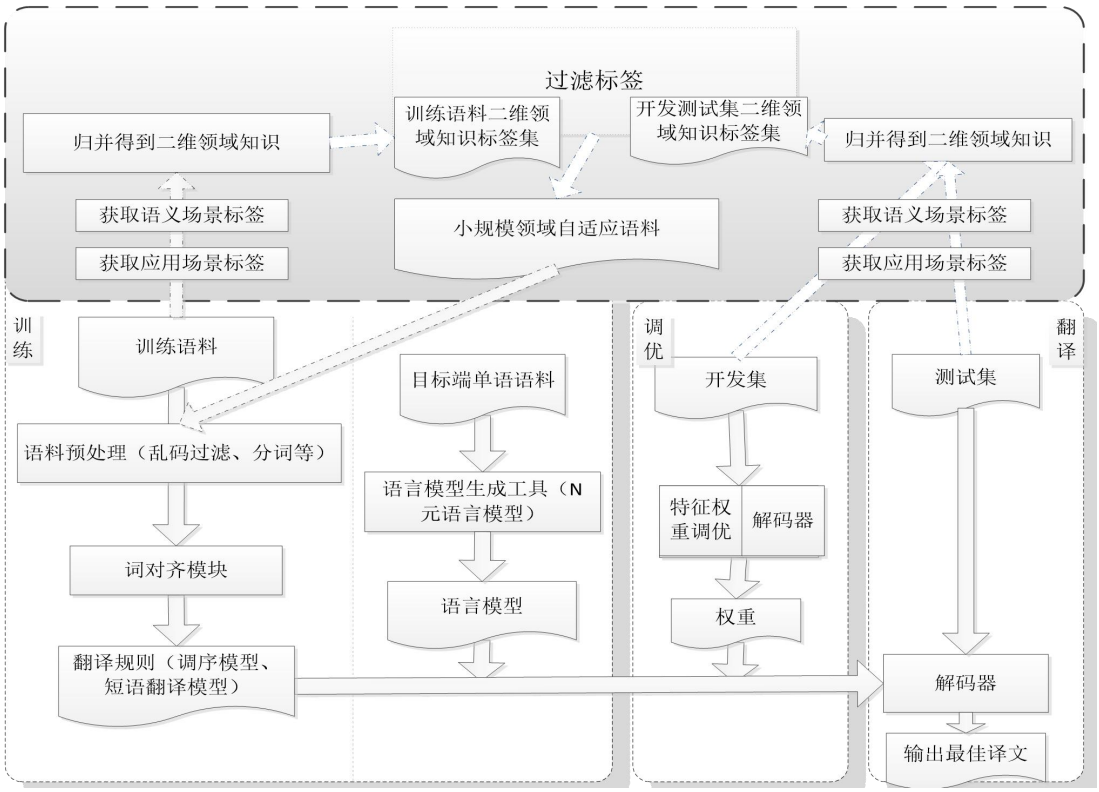

图 2 翻译系统流程图

本研究旨在对翻译系统中的训练数据按照开发集或者测试数据的二维领域知识进行筛选，以达到统计机器翻译领域自适应的目标。筛选的流程图见图 2。首先，对训练数据中每个句对中的日文词汇获取语义范畴标签和应用场景标签，并联合作为术语（或者句子）级别的二维领域标签，同样对开发集和测试集中的每个日文术语或者句子也获取二维领域标签。

表 3 为日文术语“治療薬開発（治疗药物开发）”的标注实例。将这两个维度的词汇级别的信息去重合并，按照读取先后顺序排列，得到术语(或句子)级别的二维领域知识为（抽象物；健康･医学;人工物-食べ物;科学･技術）。其后，将开发集和测试集中每个术语（或句子）所属领域的二维领域

表 3 二维领域知识标注实例

日文句子:治療薬開発		
中文句子:治疗药物开发		
分词	语义范畴	应用场景
治療	抽象物 （抽象物体）	健康･医学 （健康医学）
薬	人工物-食べ物 （人造物品-食物）	健康･医学 （健康医学）
開発	抽象物 （抽象物体）	科学･技術 （科学技术）

知识全部去重取并集，得到了目标领域的领域标签集合。在训练数据中过滤出和目标领域标签集合中的每个标签相同的句子，过滤后每个领域标签下会有不同数量的术语（或句子），按不同领域下属的句对频数进行排序，以便设置不同梯度的阈值选取语料。这样所得的训练数据中所有的术语（或句子）都与开发集和测试集在领域标签上保持了一致性。

4 实验

本节介绍实验部分，目的是测试领域标注对于面向日汉科技术语的统计机器翻译的效果影响。评价标准是大小写不敏感的 BLEU-4 打分，选用最短参考答案的长度惩罚。

基于短语的统计机器翻译系统采用东北大学自然语言处理实验室开发的 NIUTRANS[21]，所有的翻译系统参数都采用默认设置，语言模型为 3 元。实验采用的本研究组自有的日汉双语语料，从中选取了部分的训练语料、开发集和测试集。训练语料包含两部分，一部分是双语术语对，另一部分是双语句子对，开发集和测试集都是双语术语对。具体的数据统计量见表 4。

利用表 4 中的日汉实验数据集，通过 Niutrans[1]来训练 Baseline 系统，并在测试集上进行 BLEU 打分。实验采用 JUMAN[2][22]标注系统对原始训练数据、开发集与测试集中的每个日文术语（或句子）进行标注，获取每个词汇级别的语义范畴标签和应用场景标签。对各个数据集标记出的领域知识标签的统计见表 5。将开发集与测试集的分类标签取并，得到 42 个领域知识标签，形成目标领域标签集合。用其对原始训练数据句对进行过滤，过滤出 834504 条句对，将这些过滤后的训练语料按照不同标签号包含的句对数量进行排序，出现频次高的标签号包含的句对与测试语料领域相关性更强，将训练语料中这 42 个领域标签下的数据作为新的过滤语料，语料规模缩小为原来的 14.5%。

由于过滤出的语料太少，因此考虑适当放松过滤标准进行第一次过滤放松，通过对“标签-子标签”格式，诸如“人工物-その他”“人工物-金銭”“人工物-衣類”“人工物-食べ物”合并为“人工物”，规则见表 6，这样的归并方式使得语料规模缩小为原来的 24.7%。之后，将“健康･医学”、“科学･技術”合并为“科技”，将“人工物”、“自然物”、“抽象物”合并为“物体”，将“动物”、“植物”合并为“生命体”，进行第二次过滤放松，合并规则见表 6，语料规模缩小为原来的 51.46%。用过滤后的训练数据以及两次放松过滤的语料重新训练 Niutrans，与 Baseline 进行对比，翻译效果见表 7。

从表 7 可以看出，在中日科技术语的统计机器翻译中进行领域自适应后，分别用 24.7%和 51.46%的训练语料达到的 BLEU 值与用全部语料时的 BLEU 值相接近，且训练时间与计算复杂度大大减小。在第一次试验中，采用原始规模 14%的训练数据规模，达到了原始测试集 BLEU 打分 82%的翻译质量；在第二次改进试验中，采用原始规模 23.9%的训练数据规模，达到了原始测试集 BLEU 打分 95%的翻译质量；在第三次改进试验中，采用 51.46%的训练数据规模，达到了原始测试集 BLEU 打分 98.3%的翻译质量。实验结果说明利用利用日语的二维领域知识对科技术语进行领域标注，通过开发集和测试集的领域标签集合来映射到训练可以达到统计机器翻译的领域自适应的目的。

[1]http://www.nlplab.com/NiuPlan/NiuTrans.ch.html

[2]http://nlp.ist.i.kyoto-u.ac.jp/EN/index.php?JUMAN

表 4 机器翻译实验数据统计量

数据集	句对/术语	语言	句子数	平均句长	Vocabulary
训练数据	术语	日文	4269254	10	44455737
		中文	4269254	7	33708490
	句对	日文	1494874	43	65120202
		中文	1494874	33	49741866
开发集		日文	2500	5	12788
		中文	2500	6	16349
测试集		日文	2323	6	15668
		中文	2323	5	12341

表 5 领域标签数量

		二维领域知识标签数量	总句子数	无标注句子数
中日科技术语语料	原始训练集	144	5764128	669297
	开发集	35	2500	190
	测试集	36	2323	361

表 6 过滤放松规则

	合并前标签	合并后标签
第一次过滤放松	人工物-乗り物	人工物
	人工物-金銭	
	人工物-衣類	
	人工物-食べ物	
	人工物-その他	
	場所-施設部位	場所
	場所-施設	
	場所-自然	
	場所-機能	
	場所-その他	
第二次过滤放松	健康・医学	科技
	科学・技術	
	人工物	物体
	自然物	
	抽象物	
	动物	生命体
	植物	

5 总结与展望

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。本研究借鉴了统计机器翻译领域自适应方法中的数据方法和话题模型的长处，采用 JUMAN 对日文句子的领域进行领域标记，通过这个方法可以为测试数据的句子进行领域标注，通过测试数据的领域标签集合来选择训练数据，达到训练数据和测试数据的领域自适应的目的。实验表明，JUMAN 标记对句子的领域标注保证了使用部分训练数据能够达到与原始训练数据可比较的翻译效果，在几乎不损失机器翻译性能的前提下降低了翻译系统训练和解码的成本。

本研究今后需要改进的工作主要如下：

1）过滤算法仍需要提高效率，在过滤中只使用了 JUMAN 提供的标签下属句子

表 7　翻译效果

数据	系统	训练数据句对个数	语料规模（相比 baseline）	测试集 BLEU	翻译质量（相对 baseline）
中日科技术语语料	Baseline	5764128	100%	0.8012	100%
	阈值	834504	14.5%	0.6567	82%
	放松过滤Ⅰ	1421553	24.7%	0.7684	95.9%
	放松过滤Ⅱ	2966479	51.46%	0.7983	98.3%

的频次作为阈值来进行筛选。未来可以考虑利用各个领域标签之间的关系来完成过滤；

2）目前，JUMAN 分类体系中的领域信息不全，导致了一定的标引问题，且完全依赖 JUMAN 中的分类标准及其词汇间关系，对于自然语言来说这些词汇量尚显稀疏，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。

参考文献:

[1]冯志伟，现代术语学引论[M].北京：语文出版社，1997

[2]Koehn P, Och F J, Marcu D. Statistical phrase-based translation[C]//Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1. Association for Computational Linguistics, 2003: 48-54.

[3]崔磊，周明，统计机器翻译领域自适应综述，智能计算机与应用，2014 年 12 月，第 4 卷第 6 期，P31-34.

[4]Li H, Cao Y, Li C. Using bilingual web data to mine and rank translations[J]. IEEE Intelligent Systems, 2003, 18(4): 54-59.

[5]Rapp R. Automatic identification of word translations from unrelated English and German corpora[C]//Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics. Association for Computational Linguistics, 1999: 519-526.

[6]何彦青, 刘建辉, 屈鹏,等. 基于机器翻译的专利术语翻译获取方法研究[J]. 图书情报工作, 2014(19):25-30.

[7]Eck M, Vogel S, Waibel A. Low cost portability for statistical machine translation based on n-gram coverage[J]. Proceedings of Mtsummit X, 2005.

[8]Zhao B, Eck M, Vogel S. Language model adaptation for statistical machine translation with structured query models[C]//Proceedings of the 20th international conference on Computational Linguistics. Association for Computational Linguistics, 2004: 411.

[9]Lü 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.

[10]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.
[11]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.
[12]Ueffing N, Haffari G, Sarkar A. Semi-supervised model adaptation for statistical machine translation[J]. Machine translation, 2007, 21:71-94.
[13]Wu H, Wang H, Zong C. Domain adaptation for statistical machine translation with domain dictionary and monolingual corpora[C]//Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1. Association for Computational Linguistics, 2008: 993-1000.
[14]Zhao B, Xing E P. BiTAM: Bilingual topic admixture models for word alignment[C]//Proceedings of the COLING/ACL on Main conference poster sessions. Association for Computational Linguistics, 2006: 969-976.
[15]Zhao B, Xing E P. HM-BiTAM: Bilingual topic exploration, word alignment, and translation[C]//Advances in Neural Information Processing Systems. 2008: 1689-1696.
[16]Xiao X, Xiong D, Zhang M, et al. A topic similarity model for hierarchical phrase-based translation[C]//Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1. Association for Computational Linguistics, 2012: 750-758.
[17]Hashimoto C, Kurohashi S. Construction of Domain Dictionary for Fundamental Vocabulary and its Application to Automatic Blog Categorization with the Dynamic Estimation of Unknown Words' Domains[J]. Journal of Natural Language Processing, 2008, 15(5):73-97.
[18]Koehn P, Och F J, Marcu D. Statistical phrase-based translation[C]//Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1. Association for Computational Linguistics, 2003: 48-54.
[19]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.
[20]Xiong D, Liu Q, Lin S. Maximum entropy based phrase reordering model for statistical machine translation[C]//Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2006: 521-528.
[21]Xiao T, Zhu J, Zhang H, et al. NiuTrans: an open source toolkit for phrase-based and syntax-based machine translation[C]// ACL 2012 System Demonstrations. 2012:19-24.
[22]Kurohashi S, Nakamura T, Matsumoto Y, et al. Improvements of Japanese morphological analyzer JUMAN[C]//Proceedings of The International Workshop on Sharable Natural Language. 1994: 22-28.