doi:10.3772/j.issn.2095-915x.2017.03.009

扫描二维码与作者深度对话

深度学习在统计机器翻译领域自适应中的应用研究

1. 中国科学技术信息研究所 北京 100038; 2. 富媒体数字出版内容组织与知识服务重点实验室 北京 100038; 3. 北京市科学技术情报研究所 北京 100044

丁亮[1,2] 姚长青[1,2] 何彦青[1,2] 李辉[3]

摘 要 统计机器翻译往往存在待翻译文本来源多样和领域不一致的问题。为了提升面向不同领域的文本的翻译质量，需要根据待翻译文本对训练语料进行筛选以达到领域自适应的目的。目前统计机器翻译的领域自适应方法以目标数据为基准，着重利用统计技术对训练数据或者翻译模型进行领域的适应调整，缺乏明确的领域标签。本研究在本组之前研究基础上利用深度学习中卷积神经网络 (Convolutional neural network, CNN) 对短文本进行建模，构建合适的网络结构进行有监督学习，获取完整的句子语义信息，按照待翻译文本的领域信息对训练语料进行归类筛选，获取与待翻译文本领域一致的训练数据，并将其应用到统计机器翻译中。本文采用万方英文摘要在统计机器翻译系统上进行测试，仅利用部分训练数据就得到了超越原始训练数据 BLEU 打分的翻译结果，证明了本研究的有效性和可行性。

关键词： 统计机器翻译，训练语料选取，卷积神经网络，深度学习

中图分类号： G35，TP391.41

Application of Deep Learning in Statistical Machine Translation Domain Adaptation

1. Institute of Scientific and Technical Information of China, Beijing 100038, China;
2. Key Laboratory of Rich-media Knowledge Organization and Service of Diqital Publishing Content, SAPPRFT, Beijing 100038, China;
3. Beijing Institute of Science and Technology Information, Beijing 100044, China

DING Liang[1,2] YAO ChangQing[1,2] HE YanQing[1,2] LI Hui[3]

Abstract Statistical machine translation often meet problems such as the diverse sources of test data and multiple domains. In order to improve the translation quality of texts from different domains, training

基金项目： 本文受国家自然科学基金项目（61303152、71503240和71403257）和中国科学技术信息研究所重点工作项目（ZD2017-4）的资助。

作者简介： 丁亮（1994-），硕士研究生，研究方向：机器翻译，自然语言处理；姚长青（1974-），博士，副研究员，研究方向：情报理论与方法；何彦青（1974-），博士，副研究员，研究方向：机器翻译，自然语言处理，Email：heyq@istic.ac.cn；李辉（1975-），硕士，副研究员，研究方向：科技情报，危机管理，Email：lih@bjstinfo.com.cn。

corpus often needs to be filtered according to target texts to realize domain adaption. The current adaptive methods for statistical machine translation aim to the target texts and focus on the choice of training data and the adjustment of translation models. These approaches have not accuracy and explicit domain label for the texts or data. In this study, we aimed to obtain whole sentence semantic information based on our lab's pre-research. The short text was modeled by Convolutional Neural Network (CNN), and a suitable network structure was constructed for supervised learning. The training corpus was classified and selected according to the domain information of the test corpus to obtain the part training data same domain as test data. We applied this method to SMT system and test this study on the English abstracts of Wanfang data. The results showed that only part of the training data goes beyond the original training data in BLEU score. This indicated that the method is efficient and feasible.

Keywords: Statistical machine translation, training data selection, convolutional neural network, deep learning

1 引言

从 2006 年 Hinton 等人提出深度学习的概念之后，机器学习领域取得突破性的进展。深度学习是一种基于特征学习和特征层次结构的机器学习方法，它通过模拟人脑分析和学习的方式，进行逐层特征变换，每次训练一个单层网络，将样本在原空间的特征变换到一个新特征空间，从而使分类或预测更加容易。深度学习首先在语音识别和图像识别等领域获得巨大的成功，之后迅速在自然语言处理的各个研究方向上掀起热潮。

在机器翻译领域，深度学习有两类应用，其一是端到端神经机器翻译 [1]，直接利用神经网络实现源语言文本到目标语言文本的映射。其二是仍以统计机器翻译为主体框架，利用深度学习改进其中的关键模块，例如语言模型 [2,3]、调序模型 [4,5]、短语表示 [6] 等。本研究与第二种较为相似，着重于深度学习在统计机器翻译的领域自适应研究方面。

统计机器翻译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。一般来说，训练数据与目标文本的领域越接近、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。在科技文献翻译中，训练数据通常是来源繁杂，领域多样，与待翻译的目标文本的领域不能保证完全一致，为了优化目标文本的翻译成本和翻译质量，因而产生了“领域自适应”问题。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。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 [7-12]、基于混合模型的方法 [13,14]、半监督学习方法 [15,16]、话题（Topic）模型方法 [17-19] 以及基于领域标签的方法 [20,21]。前四类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

doi:10.3772/j.issn.2095-915x.2017.03.009

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。第五类方法对前四种方法的缺陷进行了很好的弥补，通过给出训练数据或者测试数据明确的领域标签，设计语料过滤算法对训练数据进行过滤，达到领域自适应的目的。

本研究在第五类方法的基础上进一步借助深度学习中卷积神经网络技术，对统计机器翻译的训练数据进行领域分类。目前基于词嵌入的卷积神经网络[22](Convolutional neural network, CNN）文本分类方法已经取得了很好的效果[23]，我们采用基于卷积神经网络的深度学习训练方法，将训练语料的单个句子看作短文本，实现了短文本分类进行领域标注，并将该技术应用到统计机器翻译中，通过对训练数据、开发集和待翻译的测试集中英语句子进行领域自动分类，可以获取句子的领域类别，进而生成测试集或训练集的领域标签集合。利用该领域标签集合筛选训练数据，从而保证领域的一致性。本研究在统计机器翻译系统上测试了该方法，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。

本文结构如下：引言之后，第二节给出统计机器翻译领域自适应的相关工作，第三节介绍了基于短语的统计机器翻译模型以及训练流程，第四节首先描述卷积神经网络的网络结构和文本分类方法，之后为统计机器翻译系统中利用类别标签对训练数据进行筛选的算法。第五节为两个实验用于验证不同领域的短文本分类的有效性，首先设计实验一验证深度学习训练的方式优于传统的统计机器学习方式（如SVM），其次设计实验二验证基于深度学习的训练数据选择对统计机器翻译系统的翻译效果的影响，第六节是总结与未来展望。

2 相关研究

统计机器翻译的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、话题模型的方法[24]，以及基于领域标签的方法五类。

基于数据选择方法是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。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]，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[10,11]，词语或者短语的覆盖率也可以用来选择训练数据[12]。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更适合在线的机器翻译，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，再适当为每个子模型分配和调整权重。Foster G等人按照训练数据的不同来源将数据进行分类[13]，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。Hal 等人的工作 [14] 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（OOV、集外词。Out Of Vocabulary），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。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。

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 采用直推式半监督学习（Transductive learning）方式 [15]。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 [16]。此类方法也没有对领域进行具体的定义。

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。一些学者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 [17,18]。这些研究都是在词级翻译上使用了话题信息。Xiao 等则通过构建层次短语翻译规则的 topic model 模型，在对测试集翻译解码时创建话题模型相似度指标，进而选取层次短语的规则 [19]。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。

基于领域标签的方法使用知识库或者机器学习算法通过对训练数据或测试数据进行明确的领域标注，设计相应的语料筛选算法，通过过滤筛选出与测试语料领域一致的较小规模训练语料，达到领域自适应的目的。丁亮等人分别使用《汉语主题词表》和日语词汇化二维知识设计语料过滤算法 [20,21]，使用较小规模训练语料得到了与原始语料可比较的翻译质量，并且大大降低了统计机器翻译的成本和系统开销。

本研究类似于第五类方法，目标在于利用神经网络方法对完整的句子建模进行句子分类，进行统计机器翻译的领域自适应。卷积神经网络是一种有特殊网络结构的神经网络，由 LeCun 等人在 1989 年提出 [25], 其变长输入、卷积、局部连接、权值共享和子采样等特点，使得 CNN 有很大的泛化能力。最初 CNN 主要应用于图片处理，随着神经语言模型的发展，文本可以很好的被表征，之后 CNN 应用于 NLP 的任务中 [26,27]，并且取得了较好的表现。Kim 基于卷积神经网络的文本分类网络结构 [9]，通过二维卷积和多通道的方式在多个语料集上训练出了比传统 SVM 模型更好的效果。同样采用 CNN 可以对短文本进行分类 [28,29]。本研究沿用 Kim 的工作，为了提升效率，将多通道简化为三通道，将全局词向量输入优化为句子输入，采用局部词嵌入方法，设计出了较为高效的端到类别（end2class）网络结构。

与已有方法相比，前四类领域自适应方法采用词典或树库等知识库作为领域信息，大都没有给出明确的既定语义标签，也没有从句子自动分类的角度进行研究。第五类方法没有考

doi:10.3772/j.issn.2095-915x.2017.03.009

虑词序，其领域标注算法大都为统计方法，句子级别的领域标签都是通过句子中词语领域标签来归并得到的，本研究把数据方法和半监督学习方法结合起来，首次采用深度学习训练短文本分类器的方法对机器翻译语料进行领域标记，统计机器翻译的语料通过深度学习分类器标引，获得每个句子的领域标签，相当于为英文句子的领域作了一种显性的表达。通过测试数据集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。

3 统计机器翻译模型和流程

3.1 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中，对于源语言句 $S=s_1s_2...s_L$ 采用对数线性模型实现翻译，过程为搜索具有最大概率的目标翻译 $T=t_1t_2...t_K$：

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

其中 $h_m(S,T)$ 为特征函数，λ_m 为特征权重。这里 phrase-based SMT 使用了如下特征函数：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。

3.2 翻译流程

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，如图 1 所示，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

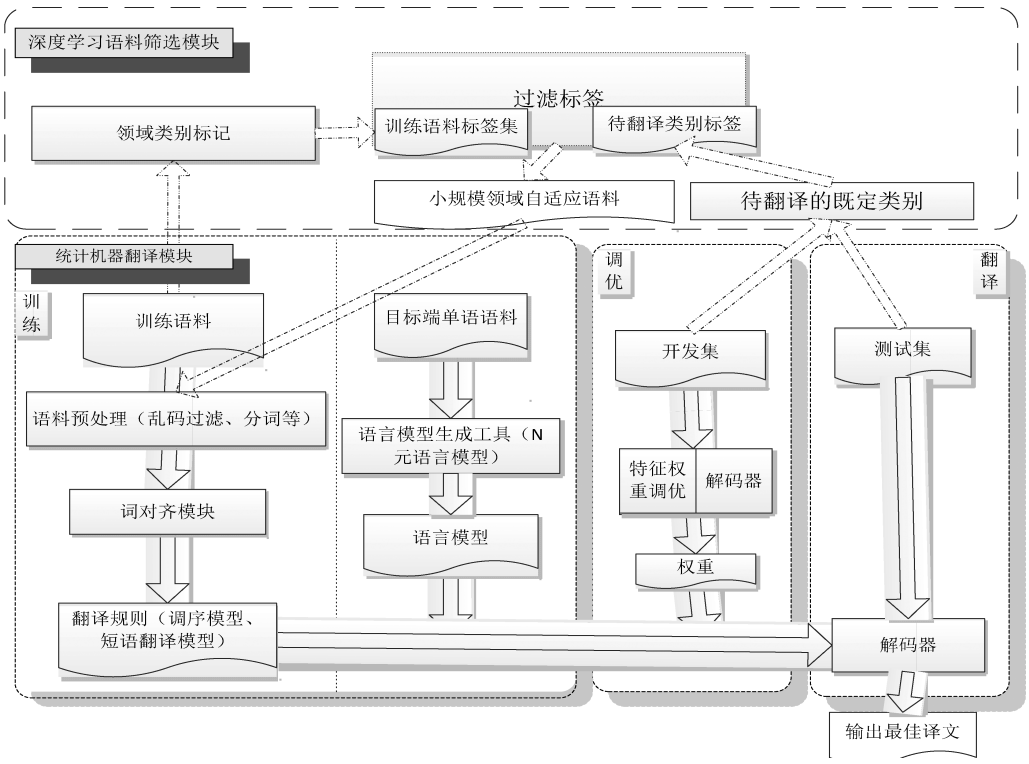

图1 翻译系统流程图

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，从而观察翻译系统的翻译效果。

4 基于深度学习的统计机器翻译训练数据选择

本节介绍利用卷积神经网络这种深度学习结构来对短文本进行半监督学习，对整个句子进行建模和特征学习，构造短文本分类器，实现基于短语的 SMT 的训练数据选择，达到领域自适应的目的。具体体现为对统计机器翻译系统中训练语料按照开发集和测试集的领域标签进行过滤，过滤的流程图见图 1。主要流程有以下几步：①利用深度学习分类器对训练语料、开发集和测试集中的英文句子进行领域类别自动标记②获取与测试数据领域类别一致的训练语料，作为小规模领域自适应语料，从而保证了领域一致性。主要模块包括构建卷积神经网络短文本分类器对句子进行标注以及训练数据选择两个模块。

4.1 基于卷积神经网络的短文本分类器

卷积神经网络作为深度学习的典型网络结构，由全连接网络改进而来，是一种改进的前馈神经网络。卷积核受到生物学上感受野（receptive field）的机制启发而提出，感受野指的是听觉系统、视觉系统中的神经机制，比如视觉神经系统中，一个神经元的感受野是神经网模上的特定区域，只有该区域受到刺激，才会激活该特定的神经元。相比传统全连接网络结构（图 2 左），卷积神经网络（图 2 右）可以极大的减少训练复杂度，增强局部感知能力。卷积神经网络有局部连接、权重共享和子采样这样几个特点，卷积神经网络将文本词向量经过多次卷积处理得到更抽象的特征，这种多次抽象的训练过程使得计算机获取更能代表类别特征的向量，在最后一层用 softmax 层 $\sigma(z)_j = \frac{e^{z_j}}{\sum_{k=1}^{K} e^{z_k}}$ 实现分类。

本研究中设计的卷积神经网络结构如图 3 所示，首先是句子输入查表窗口（look-up table），给定固定长度 n 的输入窗口，通过简单的 one hot 映射，生成如 [0，1，0，0，0，0] 的词表示，设置 128 维的词嵌入维度进行 word2vec 训练，进行高维词向量映射将训练后的词向量作为卷积神经网络的初始输入。对于

doi:10.3772/j.issn.2095-915x.2017.03.009

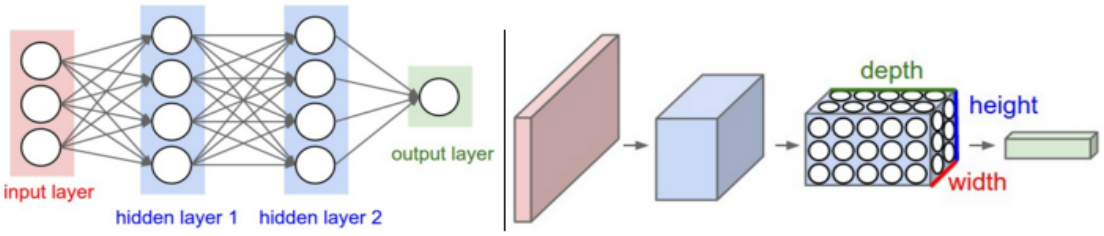

图2 用于分类的Softmax层

卷积神经网络的输入层，将第 i 个词的 k 维度词向量表示为 x_i，则对于长度为 n 的句子可以表征为 $x_{1:n}=x_1\oplus x_2\oplus...\oplus x_n$，其中⊕是连接符。在多通道滑动卷积这一步，采用三个 k*r 的过滤器（filter）构成三个不同的通道，以便构造多特征。对于每个过滤器，将 $x_{i:i+h-1}$ 这 h 个词向量作为窗口构造新特征 $c_i=f(w*x_{i:i+h-1}+b)$，其中卷积核 f 是非线性函数（此处我们采用反正切函数 tanh(θ)），并将这种过滤窗口滑动计算整个长度为 n 的句子，得到长度为 h–1 的特征向量 $C^i=[c_1,c_2,\ldots,c_{n-h+1}]$，在本研究中滤波器长度分别为 3、4、5 的过滤器，则生成的多通道卷积特征向量有 C^1、C^2、C^3。

将多通道卷积神经网络生成的特征向量 C^1、C^2、C^3 进行池化子采样（pooling）操作，池化子采样可以提取出表征向量中最有代表性的子向量，在研究中具体采用 1-max-pooling 作为特征提取函数，得到每个通道的子采样特征 $\hat{c}^i=\max\{c_1^i,c_2^i,...,c_{n-h+1}^i\}$。

将子采样特征 $\hat{c}^1$、$\hat{c}^2$、$\hat{c}^3$ 输入最后一层分类层，在本研究中采用 Softmax 作为分类函数，Softmax 模型 $\sigma(z)_j=\frac{e^{z_j}}{\sum_{k=1}^{K}e^{z_k}}$ 是 logistic 模型在多分类问题上的推广，将输入变量进行归一化操作，并重新分配每个维度的值，便于分类。

为了防止样本小造成的模型参数过拟合，采用 Hinton 等人于 2012 年提出的 Dropout 方法 [30]，在模型训练时随机让网络某些隐含层节点的权重不工作，不工作的那些节点可以暂时认为不是网络结构的一部分，但是它的权重得保留下来，使得下次样本输入时它又能继续工作。

在本研究中可采用卷积神经网络的模型训练深度学习短文本分类器，可以为语料标记类别标签。

4.2 基于分类标签筛选训练数据

首先，基于深度学习短文本分类器对训练数据中每个句对中的英文句子进行领域标记，获取每个英文句子的类别标签，同样按照既定的开发集和测试集中的英文句子的领域类别，用来选择与其有相同类别标签的训练数据。例如训练语料中句子“extensive simulations show the efficiency of the algorithm”经深度学习短文本分类器的自动判别后，得到类别标签为 TP（自动化计算机），表示该句子为自动化计算机类别，其中 TP 为中图分类号的二级目录，代表了该句子的领域信息。其后，根据开发集和待翻译测试集的已知领域信息在训练数据中过滤出和测试集领域类别相同的句子。过滤之后生成训练集的子集，这样所得的训练数据中所有的句子都与开发集测试集在领域标签上保持了一致性。

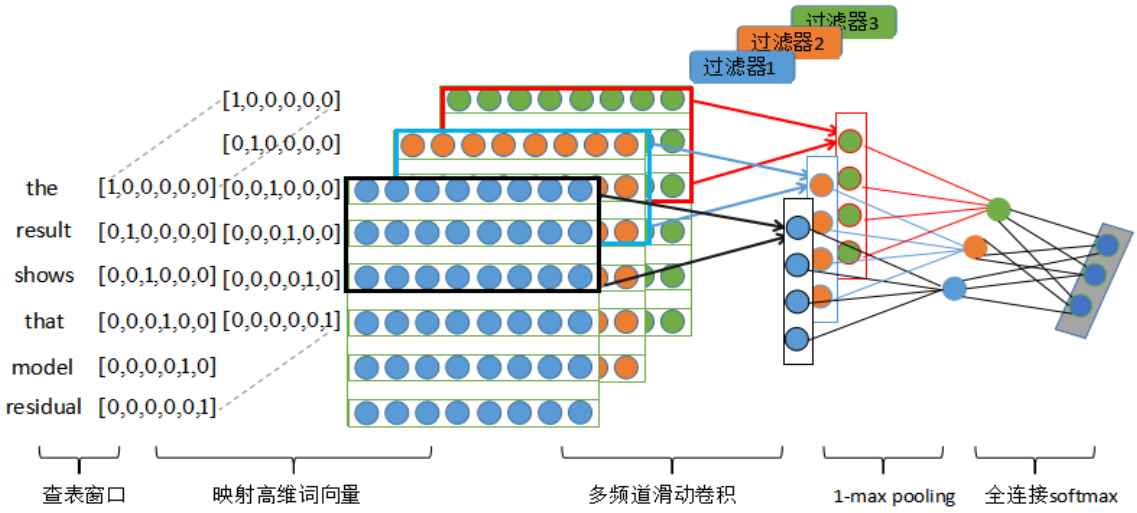

图3 基于卷积神经网络的深度学习网络结构图

5 实验

实验研究目标有两个，一是验证深度学习短文本分类器对英文句子分类效果，另一是验证训练数据类别标注和数据筛选对于英汉统计机器翻译效果影响。前者评价标准是分类正确率，即测试集中分类正确样本个数比上测试集样本容量；后者评价标准是大小写不敏感的 BLEU-4 打分，选用最短参考答案长度惩罚。

5.1 深度学习短文本分类器的领域分类效果

本实验将对比深度学习短文本分类器与机器学习中表现最佳的支持向量机 SVM 分类器 [31]，实验分别在表 1 的五个领域各选出了 4500 和 500 条英文短句子分别作为训练集和测试集，分别进行了二分类和多分类实验，所选的五个领域有四个为工业类下属的相近领域（TD 矿业工程、TH 机械仪器、TP 自动化计算机、TU 建筑），另外一类为领域相差较远的（A 马克思）。利用 4.1 节设计的卷积神经网络结构，词嵌入的维度设置为 128，三组过滤器的高度设置为 3、4、5，dropout 的比例设置为 0.5。每一组实验我们均采用测试集中分类正确的句子除以测试集总数作为分类正确率。训练数据和测试数据的统计量见表 1。

表1 五个领域的英文句子统计量

所属领域		TD 矿业工程	TH 机械仪表	TP 自动化计算机	TU 建筑	A 马克思
训练集	句子数	4500	4500	4500	4500	4500
	平均句长	36	33	31	34	32
	Vocabulary	163901	151603	143901	156915	145908

在本研究中，我们为了验证所设计的深度学习网络结构对于二分类问题的高度准确性，

doi:10.3772/j.issn.2095-915x.2017.03.009

我们设计了五组分类实验，其中每一组实验都用 SVM 分类器作为基准 baseline。第一组二分类对照实验采用领域相近的 TD 矿业工程和 TH 机械仪表的训练集共 9000 句，测试集采用 TD 和 TH 的测试集共 1000 句；第二组二分类对照实验同样采用领域相近的 TP 自动化计算机和 TU 建筑的训练集共 9000 句，测试集采用 TP 和 TU 的测试集共 1000 句；第三组二分类对照实验仍然采用领域相近的 TD 矿业工程和 TU 建筑的训练集共 9000 句，测试集采用 TD 和 TU 的测试集共 1000 句；第四组二分类对照实验我们采用领域差别较大的 A 马克思和 TH 机械仪表的训练集共 9000 句，测试集采用 A 和 TH 的测试集共 1000 句；第五组多分类对照试验我们采用领域相近的 TD 矿业工程、TH 机械仪表、TP 自动化计算机和 TU 建筑类的训练集共 18000 句，测试集采用这四个类别对应的测试集共 2000 句。实验结果见表 2。

表2 三组对照试验的统计与结果

实验	语料统计		分类正确率（%）	
	训练集	测试集	深度学习分类器	SVM 分类器
实验一	TD+TH 共 9000	TD+TH 共 1000	98.76%	85.30%
实验二	TP+TU 共 9000	TP+TU 共 1000	98.97%	83.14%
实验三	TD+TU 共 9000	TD+TU 共 1000	99.39%	81.64%
实验四	A+TH 共 9000	A+TH 共 1000	99.91%	92.60%
实验五	TD+TH+TP+TU 共 18000	TD+TH+TP+TU 共 2000	51.38%	75.95%

备注（TD: 矿业工程、TH：机械仪器、TP：自动化计算机、TU：建筑）

如图 4 通过前四个对照试验，我们发现传统的 SVM 分类器对于领域差别较大的问题（实验 4 中 A 马克思和 TH 机械仪器）分类结果最好，但是对于领域差别较小的问题（前三个实验都是工业类 :TD 矿业工程、TH 机械仪器、TP 自动化计算机、TU 建筑）分类结果明显下降；深度学习分类器对于领域差别不一的任务几乎都达到百分百的正确率。充分证明了本研究所设计的深度学习分类器对于二分类问题的有效性。通过对照实验 5，我们发现对于四个类别的相似领域多分类问题，SVM 和深度学习分类器的分类效果都不理想。通过多组对照实验我们发现，我们设计的深度学习分类器对于多分类问题不敏感，但是对于二值分类问题效果显著，明显优于传统统计分类方法。本实验验证了我们设计的深度学习短文本分类器对于二分类问题的有效性。

5.2 机器翻译效果评价

在本研究中我们的 phrase-based SMT 系统采用 NIUTRANS[32]，参数采用默认，语言模型采用 3 元文法。实验采用本研究组自有的万方论文摘要数据集，选“T 工业”大类下的 TD 矿业工程、TH 机械仪器、TP 自动化计算机和 TU 建筑各 20W 条句对，混合四个类别后的训练语料共计 80W 条句对，开发集从 TP 领域选取 4000 条平行语料，测试集从 TP 领域选取 2000 条平行语料。具体的数据统计量见表 3。

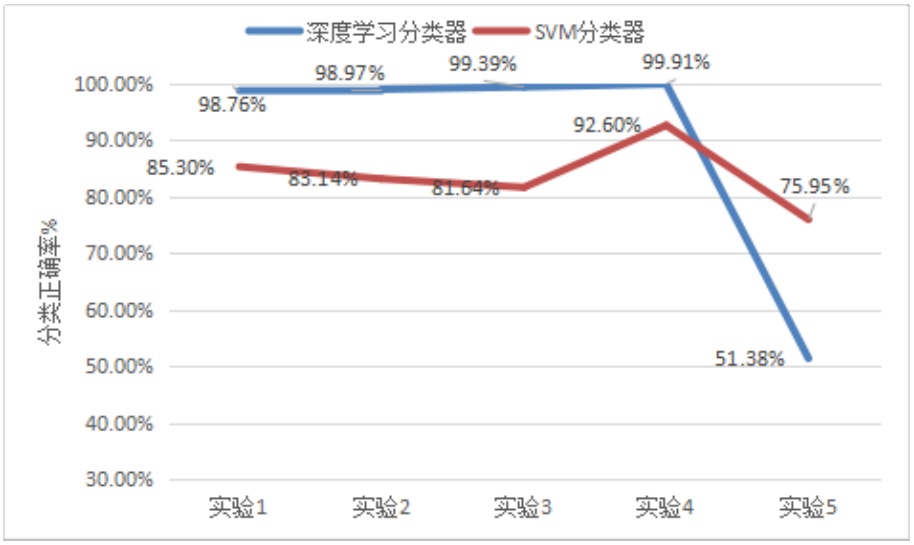

图4 三组对照实验标记正确率

表3 机器翻译实验数据统计量

数据来源	数据集	语言	句子数	平均句长	Vocabulary
万方数据	训练数据	英文	800000	29	23275889
		中文	800000	32	25693249
	开发集	英文	4000	29	116267
		中文	4000	32	128319
	测试集	英文	2000	28	56146
		中文	2000	33	66305

利用表 3 中的实验数据集，通过 Niutrans 来训练 Baseline 系统，并在测试集上进行 BLEU 打分。然后用基于卷积神经网络的深度学习短文本分类器对本研究组所有的 T 工业大类数据集中 TP 和非 TP 两类英文句子进行训练，对 80W 训练数据进行类别标注，其中，TP（True Positive）表示被判定为正样本，事实上也是正样本；FP（False Positive）表示被判定为正样本，但事实上是负样本；FN（False Negative）表示被判定为负样本，但事实上是正样本；TN（True Negative）表示被判定为负样本，事实上也是负样本。用分类号作为每个句子的类别标签。用深度学习短文本分类器和传统 SVM 分类器对万方数据训练数据标记的效果统计见表 4. 我们发现采用深度学习方法进行领域自适应可以使得语料规模缩小到原来的 30.09%，且保证 99.38% 的分类正确率，再次证明了对传统统计方法方法的突破。

表4 机器翻译训练语料筛选情况

短文本分类方法	训练语料					
	原始规模	筛选后规模	规模缩小比率	包含 TP 类句子	TP 正确分类条数	分类正确率
传统 SVM	800000	326001	40.75%	200000	137000	68.50%
深度学习	800000	240756	30.09%	200000	198700	99.35%

doi:10.3772/j.issn.2095-915x.2017.03.009

本研究分别使用传统 SVM 方法筛选的 326001 对小规模平行语料和深度学习方法筛选的 240756 对小规模平行语料作为新的训练集，并用于训练两个新的翻译系统与 Baseline 进行对比，并在测试集上进行 BLEU 打分，其翻译效果见表 5。可以看出，采用传统 SVM 方法在万方论文摘要数据集上进行领域自适应之后，仅用原始规模 40.75% 的训练语料可以得到高出基准系统 0.08 百分点的 BLEU 打分；采用深度学习方法在万方论文摘要数据集上进行领域自适应后，仅用原始规模 30.09% 的训练语料可以得到高出基准系统 0.13 百分点的 BLEU 打分。实验结果说明利用深度学习短文本分类器对英文的领域标注，筛选出与测试集领域一致的训练数据可以达到统计机器翻译领域自适应的目的。

表5 万方摘要数据翻译效果

数据	系统	训练数据句对数	测试集 BLEU（百分点）
万方数据	Baseline	800000	20.41
	传统 SVM 过滤	326001	20.49
	深度学习过滤	240756	20.54

6 结论及未来改进方向

统计机器翻译常常面临训练数据与待翻译测试集句对类别领域不一的问题，这种领域不一致不仅增加了翻译成本也影响了翻译性能，因此统计机器翻译领域自适应问题一直是本行业致力解决的问题。本研究借鉴了统计机器翻译领域自适应方法中数据方法和半监督学习方法，首次采用深度学习训练短文本分类器的方法对机器翻译语料进行领域标记，通过深度学习分类器标引获得每个句子的领域标签，相当于为英文句子的领域作了一种显性的表达。通过测试数据领域来选择训练数据，达到训练数据和测试数据领域自适应的目的。实验表明，采用深度学习短文本领域分类的方法明显优于传统 SVM 之类的方法，且仅仅使用部分训练数据就达到了比原始训练数据效果更好的翻译效果，在提升翻译性能的同时降低了翻译系统训练和解码成本。

本研究需要改进之处主要如下：1）我们仅仅使用了基于卷积神经网络的深度学习网络结构，在未来的研究中我们可以考虑融合循环神经网络 (RNN, Recurrent Neural Networks）更好的增强短文本分类器分类任务适应性。使其不仅仅在二分类上表现优异，对于多分类问题也可以有很好的表现。2）在深度学习中引入现有知识库，改进深度学习的优化函数，增加知识库判别权重，这也将提升深度学习领域辨别技术在统计机器翻译上的效率和精度。

参考文献

[1] Sutskever I, Vinyals O, Le Q V. Sequence to Sequence Learning with Neural Networks[J]. Advances in Neural Information Processing Systems, 2014(4): 3104-3112.

[2] Mikolov T, Chen K, Corrado G, et al. Efficient Estimation of Word Representations in Vector Space[J]. Computer Science, 2013.

[3] Vaswani A, Zhao Y, Fossum V, et al. Decoding with Large-Scale Neural Language Models Improves Translation[C]// EMNLP. 2013: 1387-1392.

[4] Auli M, Galley M, Quirk C, et al. Joint Language and Translation Modeling with Recurrent Neural Networks[J]. American Journal of Psychoanalysis, 2013, 74(2): 1044-1054.

[5] Li P, Liu Y, Sun M. Recursive Autoencoders for ITG-Based Translation[C]// EMNLP. 2013: 567-577.

[6] Zhang J, Liu S, Li M, et al. Bilingually-constrained Phrase Embeddings for Machine Translation[C]// Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, Baltimore, Maryland, USA, June 23-25, 2014: 111–121.

[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, the University of Geneva, Switzerland. 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] Moore R C, Lewis W. Intelligent Selection of Language Model Training Data[C]// ACL 2010, Proceedings of the, Meeting of the Association for Computational Linguistics, July 11-16, 2010, Uppsala, Sweden, Short Papers. 2010:220-224.

[11] Axelrod A, He X, Gao J. Domain Adaptation via Pseudo in-domain Data Selection[C]// Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Edinburgh, UK. 2011: 355-362.

[12] 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.

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

[14] Daum, Iii H, Jagarlamudi J. Domain Adaptation for Machine Translation by Mining Unseen Words[C]// Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers. Association for Computational Linguistics, 2011: 407-412.

[15] Ueffing N, Haffari G, Sarkar A. Semi-supervised Model Adaptation for Statistical Machine Translation[J]. Machine Translation, 2007, 21(2): 77-94.

[16] Wu H, Wang H, Zong C. Domain Adaptation for Statistical Machine Translation with Domain Dictionary and Monolingual Corpora[C]// International Conference on Computational Linguistics. Association for Computational Linguistics, 2008: 993-1000.

[17] Zhao B, Xing E P. BiTAM: Bilingual Topic AdMixture Models for Word Alignment[C]// Coling/ACL on Main Conference Poster Sessions. Association for Computational Linguistics, 2006: 969-976.

[18] Zhao B, Xing E P. HM-BiTAM: Bilingual Topic Exploration, Word Alignment, and Translation[C]// Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December. DBLP, 2007:1689-1696.

[19] Xiao X, Xiong D, Zhang M, et al. A Topic Similarity Model for Hierarchical Phrase-based Translation[C]// Meeting of the Association for Computational Linguistics: Long Papers. Association for Computational Linguistics, 2012: 750-758.

doi:10.3772/j.issn.2095-915x.2017.03.009

[20] 丁亮 , 李颖 , 何彦青 , 等 . 基于汉语主题词表的统计机器翻译训练数据筛选方法及实验研究 [J]. 情报学报 , 2016, 35(8): 875-884.

[21] 丁亮 , 李颖 , 何彦青 , 等 . 基于二维词汇化领域知识的日汉科技术语翻译方法研究 [C]. CWMT 2016, 第十二届全国机器翻译研讨会 (CWMT 2016). 2016, 论文集 : 19-28.

[22] LeCun Y, Bengio Y, Hinton G. Deep Learning[J]. Nature, 2015, 521(7553): 436-444.

[23] Kim Y. Convolutional Neural Networks for Sentence Classification[J]. Eprint Arxiv, 2014.

[24] 崔磊 , 周明 . 统计机器翻译领域自适应综述 [J]. 智能计算机与应用 , 2014, 4(6): 31-34.

[25] Lecun Y, Boser B, Denker J S, et al. Backpropagation Applied to Handwritten Zip Code Recognition[J]. Neural Computation, 1989, 1(4): 541-551.

[26] Collobert R. Natural Language Processing From Scratch[J]. The Journal of Machine Learning Research, 2011.

[27] Yih W T, He X, Meek C. Semantic Parsing for Single-Relation Question Answering[C]// Meeting of the Association for Computational Linguistics. 2014: 643-648.

[28] Kalchbrenner N, Grefenstette E, Blunsom P. A Convolutional Neural Network for Modelling Sentences[J]. Eprint Arxiv, 2014, 1.

[29] Ji Y L, Dernoncourt F. Sequential Short-Text Classification with Recurrent and Convolutional Neural Networks[C]// Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2016: 515-520.

[30] Hinton G E, Srivastava N, Krizhevsky A, et al. Improving Neural Networks by Preventing Co-adaptation of Feature Detectors[J]. Computer Science, 2012, 3(4): 212-223.

[31] Yu H F, Ho C H, Arunachalam P, et al. Product Title Classification versus Text Classification[EB/OL]. [2017-01-12]. http://ntucsu.csie.ntu.edu.tw/~cjlin/papers/title.pdf.

[32] 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.