doi:10.6043/j.issn.0438-0479.202005002

文化风格区分的无监督领域适应的电商产品翻译

史小静,宁秋怡,段湘煜*
(苏州大学计算机科学与技术学院,江苏 苏州 215006)

摘要:电商产品翻译系统的训练存在两个主要的问题:电商领域训练数据稀缺和电商产品描述文化风格差异较大.为此,通过获取大量的电商产品数据信息作为训练语料,并利用基于无监督领域适应的混合训练和文化风格区分的方法改善电商产品翻译系统的性能.具体地:一方面将基于外领域数据训练得到的翻译系统应用于电商领域单语数据得到伪平行语料,使用伪平行语料进行混合训练进一步得到新的模型;另一方面给不同语言的电商数据添加对应的文化风格区分标记,在训练过程中告诉模型当前数据的所属类别,根据类别信息获取相应的文化风格区分特征向量,从而提高电商领域产品信息翻译的准确度.实验结果表明,该方法优于多种基于单语语料的电商产品翻译方法.

关键词:机器翻译;领域适应;无监督
中图分类号:TP 391.2 **文献标志码**:A **文章编号**:0438-0479(2021)06-1011-08

经济全球化促进了世界多边贸易体制的形成,跨境贸易显得日趋重要,出口产品信息翻译需求也日益突显,仅依靠人工来实现翻译花费较大且不易实现.机器翻译(machine translation, MT)的发展使得这一问题得到了缓解.近年来神经机器翻译[1-3](neural machine translation,NMT)的提出极大地改善了 MT 模型的性能,并且在某些领域已经达到了较高的水准[4-7],但是 MT 系统的性能很大程度上依赖于平行语料的规模和质量,然而在跨语言的信息处理任务中,平行语料是非常稀有的,因此基于领域适应的 MT 引起了广泛的关注,其中大部分工作集中在少量内领域平行数据可用的情况[8-9].

现有的无监督领域适应的 NMT 主要通过生成内领域伪平行数据或对模型结构进行修改. Freitag 等[10]提出使用外领域的平行语料预训练一个翻译系统,基于外领域翻译模型,利用内领域的数据继续进行调参训练,以达到领域适应的目的,同时保证了外领域系统的性能仅有稍微的减弱. Sennrich 等[11]提出将内领域目标语言文本通过反向翻译的方法翻译为源端语言数据,再将翻译得到的源端数据和真实的目标端数据构建内领域的伪平行数据. Currey 等[12]提出将内领域的目标语言文本复制到源语言端来创建内领域的伪平行语料数据. Zeng 等[13]提出分别使用外领域和内领域的数据进行预训练,然后基于知识蒸馏迭代地执行双向知识转移帮助模型的训练. Chu 等[14]提出把多语言翻译和领域自适应结合起来,改善资源缺乏的内领域的翻译模型的性能. Dou 等[15]提出将领域特征嵌入到神经网络中编码端的无监督领域适应方法,并通过多任务学习来联合训练整个神经网络. Yang 等[16]提出通过引入两个分类器,其中一个用于判断模型生成的句子是否属于目标端领域,另一个用于判断译文句子是否属于源端句子领域,这两个分类器构成了一个对抗训练的网络. Su 等[17]提出采用多任务学习的方式,将翻译任务和领域分类任务进行联合建模:通过在编码端引入领域分类器和对抗领域分类器对输入句子进行领域分类,从而分离出领域专有信息和领域共享信息;解码端使用基于注意力机制的领域分类器,从而使分类器导出的注意力权重具有领域特征,可用来调整训练过程中反馈的误差. Zeng[18]提出将 NMT 模型和单语领域分类任务联合,使用两个门控向量构建领域区分和领域共享的注释,利用目标端领域分类器得到的注意力信息调整目标词的权

收稿日期:2020-05-06 **录用日期**:2020-09-25
基金项目:国家自然科学基金(61673289)
* **通信作者**:xiangyuduan@suda.edu.cn
引文格式:史小静,宁秋怡,段湘煜.文化风格区分的无监督领域适应的电商产品翻译[J].厦门大学学报(自然科学版),2021,60(6):1011-1018.
Citation:SHI X J,NING Q Y,DUAN X Y. Culture-style aware e-commerce product translation based on unsupervised domain adaptation[J]. J Xiamen Univ Nat Sci,2021,60(6):1011-1018. (in Chinese)

重,使得领域相关的词获得更大的权重. Shoetsu 等[19]提出词表自适应方法,在微调之前将词嵌入映射到内领域的词嵌入空间,缓解领域差异较大的预训练导致的领域不匹配问题. Gordonm 等[20]提出将知识蒸馏和领域适应相结合,提升多语言对模型的效果.

NMT 模型的性能很大程度上依赖于训练数据的数量和质量,然而据本文调研,目前电商领域还没有公开可利用的平行语料,这是训练电商领域产品翻译系统的主要挑战之一. 此外,由于不同地区的文化风格和语言特点的差异,即使对于同一种产品也会有不同风格的描述信息,这是电商领域产品信息翻译的另一难点. 为了解决电商领域语料稀少这一问题,本文分别从不同的电商平台获取了不同语言的产品数据信息,主要包括中文和英文电商领域产品的数据信息,中文电商领域的语料数据取自淘宝官方网站,英文电商领域的数据语料取自亚马逊官方网站. 针对产品信息文化风格差异这一问题,本文提出了基于无监督领域适应的混合训练方法和文化风格区分方法. 利用资源丰富的新闻领域的平行语料训练源语言到目标语言以及目标语言到源语言的两个翻译系统,然后对电商领域的单语数据进行翻译得到伪的平行数据,使用伪平行数据进行混合训练和文化风格区分的方法进行模型训练.

1 混合训练和文化风格区分的方法

本文提出的混合训练和文化风格区分的无监督领域适应电商产品信息的翻译方法,使得基于资源丰富的外领域的平行语料库训练的翻译模型能够适应于没有平行语料的电商领域单语数据的翻译任务,提升电商领域的 MT 译文质量. 本文基于目前效果最好的 Transformer 进行混合训练和文化风格区分实验,将电商领域的单语数据视为内领域的数据,将新闻领域的数据视为外领域的数据.

1.1 混合训练方法

Edunov 等[21]发现利用反向翻译得到的数据集训练的模型可以提高目标端数据的测试集的效果,Bogoychev 等[22]发现使用前向翻译得到的数据集训练的模型能够提高源端数据的测试集的效果. 结合两者的论证结果和目标端复制方法,本文提出混合训练的方法,利用电商内领域的单语数据,实现无监督领域适应电商产品信息的翻译;该方法不改变 Transformer 模型的结构,是对单语数据的使用方法的创新. 首先利用新闻领域平行语料训练两个基础的 Transformer 翻译模型,对电商领域的中文商品数据信息和英文商品数据信息分别进行翻译,得到对应的伪平行数据$\{C_{\mathrm{tb}}, E'_{\mathrm{tb}}\}$和$\{E_{\mathrm{am}}, C'_{\mathrm{am}}\}$,其中,$C_{\mathrm{tb}}$和$E'_{\mathrm{tb}}$分别表示淘宝中文真实数据和英文伪数据,$E_{\mathrm{am}}$和$C'_{\mathrm{am}}$分别表示亚马逊英文真实数据和中文伪数据. 然后将$C_{\mathrm{tb}}$和$C'_{\mathrm{am}}$进行混合得到$\{C_{\mathrm{tb}}, C'_{\mathrm{am}}\}$,与之对应将$E'_{\mathrm{tb}}$和$E_{\mathrm{am}}$进行混合得到$\{E'_{\mathrm{tb}}, E_{\mathrm{am}}\}$,得到不同电商平台的单语数据. 将两组伪平行数据进行合并,使用混合后的伪平行数据分别训练中英和英中 Transformer 翻译模型,再利用新的 Transformer 翻译模型对电商领域的单语数据进行翻译,构成新的伪平行数据;然后将之混合,重新训练新的 Transformer 翻译模型;如此重复,直至中英和英中两个方向的 NMT 系统在电商领域测试集上的效果均达到最优. 本文将这种仅使用内领域不同平台的伪平行数据训练领域适应的 NMT 系统的方法称作混合训练.

1.2 文化风格区分的电商产品翻译方法

电商领域中,不同语言的不同电子商务平台的产品信息描述表现出显著的风格差异,例如给出的同一类产品,不同语言的电子商务平台的相应特性描述如下:

淘宝平台:阿迪达斯 adidas 男鞋 女鞋 2021 春季 中底 运动鞋 减震 跑步鞋

亚马逊平台:These adidas running shoes are designed to turbo charge your daily miles. A soft, comfortable elastane heel allows for natural movement of the Achilles.

从上述样例中可以看出:中文淘宝平台的产品描述信息主要是以词汇的无序堆叠方式呈现,包含较少的语义信息;与之相比,亚马逊平台的英文产品描述信息更加流畅自然并且语义信息较为丰富. 为了区分不同语言数据的不同文化风格,缓解电商产品翻译过程中的文化风格差异问题,本文给不同语言的电商数据添加了对应的文化风格区分标记(如图 1 所示). 在训练过程中告诉模型当前数据的所属类别,根据类别信息获取相应的文化风格区分特征向量. 在编码端,网络的输入信息添加源语言端的文化风格特征向量,而在解码端添加目标端语言的风格特征向量$\boldsymbol{\theta}_{\mathrm{culture}}$,使得模型在解码过程中能够学习到特定的目标语言的文化风格,在忠于源端数据的前提下使得到的译文风格更趋于目标端的风格. 文化风格特征向量在模型训练过程中与其他参数共同训练,通过与 1.1 节的混合训练方法相结合,使得模型不仅能够学习到同一产品的相关联的描述信息,同时也能够捕获到同一产品的

不同文化风格的描述信息,共同提升产品翻译的译文质量.

具体地,本文中文化风格区分的特征向量$\boldsymbol{\theta}^{\mathrm{tb}}_{\mathrm{culture}}$和$\boldsymbol{\theta}^{\mathrm{am}}_{\mathrm{culture}}$分别用来区分标记淘宝电商平台数据和亚马逊电商平台数据文化风格特征,将网络中每一层的输入和其对应的文化风格特征向量拼接在一起,共同作为下一层网络的输入.新增的文化风格区分的特征向量与网络中其他参数一起训练.文化风格区分的特征向量分别在编码端和解码端的每一层添加,并且和隐藏层的状态向量维度保持一致.

当编码端的数据为淘宝平台的数据时,则在编码端添加淘宝数据的文化风格特征向量.相应地,希望目标端的译文在忠于原文的前提下具有亚马逊文化风格的特征,则在解码端添加亚马逊数据的文化风格特征向量.具体如式(1)和(2)所示.

$$\boldsymbol{I}^{\mathrm{en}}_i = \boldsymbol{O}^{\mathrm{en}}_{i-1} + \boldsymbol{\theta}^{\mathrm{tb}}_{\mathrm{culture}}, \tag{1}$$

$$\boldsymbol{I}^{\mathrm{de}}_i = \boldsymbol{O}^{\mathrm{de}}_{i-1} + \boldsymbol{\theta}^{\mathrm{am}}_{\mathrm{culture}}, \tag{2}$$

其中,$\boldsymbol{I}^{\mathrm{en}}_i$ 和 $\boldsymbol{I}^{\mathrm{de}}_i$ 分别表示编码端和解码端第 i 层网络的输入信息,$\boldsymbol{O}^{\mathrm{en}}_{i-1}$ 和 $\boldsymbol{O}^{\mathrm{de}}_{i-1}$ 分别表示编码端和解码端第 $i-1$ 层网络的输出信息,$\boldsymbol{\theta}^{\mathrm{tb}}_{\mathrm{culture}}$ 和$\boldsymbol{\theta}^{\mathrm{am}}_{\mathrm{culture}}$ 分别表示淘宝数据的和亚马逊数据的文化风格特征向量.

相应地,混合训练时当编码端的数据为亚马逊平台的数据时,则在编码端添加亚马逊数据的文化风格特征向量.此时,希望目标端译文在忠于原文数据的前提下具有淘宝文化风格的特征,则在解码端添加淘宝数据的文化风格特征向量.特别地,文化风格区分的特征向量分别在编码器端和解码器端的每一层均添加,为了方便拼接,特征向量的维度和隐藏层的状态向量维度保持一致.

2　对比训练模型和策略

本文选择标准的 Transformer[7] 模型结合混合训练和文化风格区分的方法进行训练,并将本文提出的方法与如下几种训练策略进行对比:

a) 反向翻译　Sennrich 等[11] 提出利用外领域已

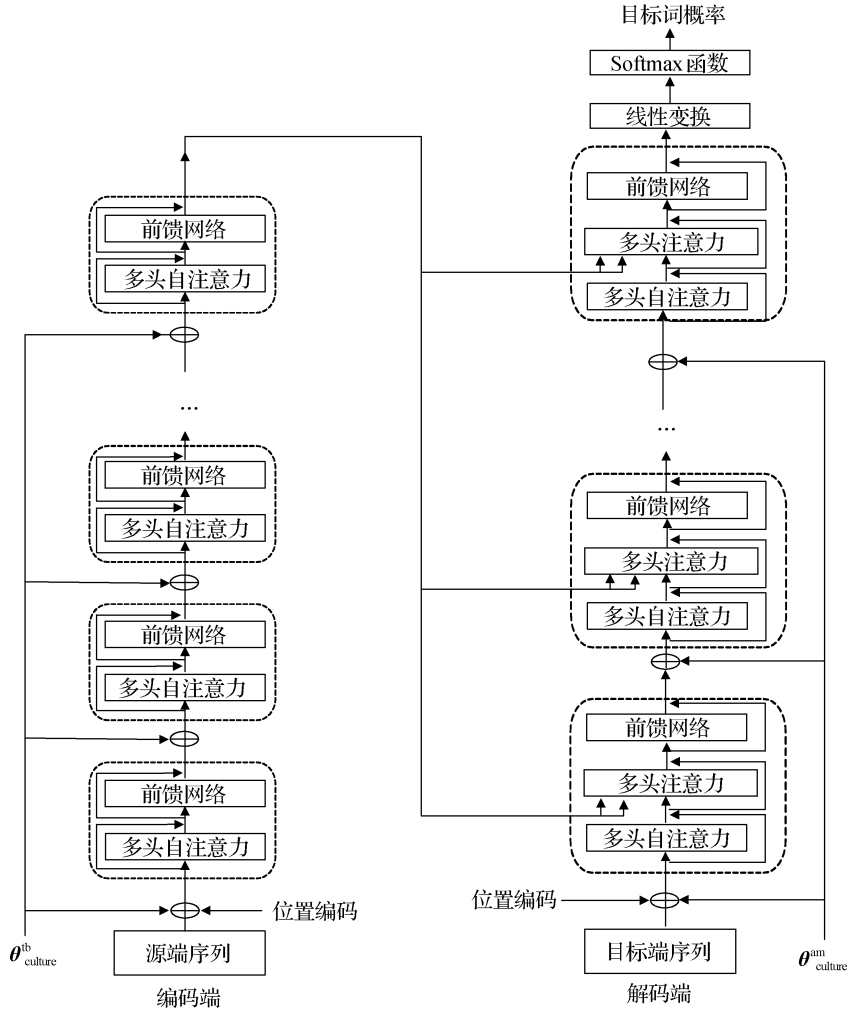

图 1　文化风格区分的网络结构

Fig. 1　The network structure of cultural-style aware

有的平行语料,训练一个目标端到源端的翻译系统,再通过训练好的系统将目标端单语语料翻译成源端对应的语料,将得到的源端语料和真实的目标端语料构成伪平行语料.将外领域的平行语料和合成的内领域的伪平行语料连接实现对内领域数据的扩充,使用扩充的语料训练源端到目标端的 NMT 系统.

b) 基于外领域模型微调的反向翻译 Freitag 等[10]提出使用外领域的平行语料预训练一个 Transformer 翻译系统,基于外领域数据训练得到的翻译模型利用内领域的数据继续进行调参训练,以达到领域适应的目标,同时保证了外领域系统的翻译性能仅有稍微的减弱.

c) 目标端复制 Currey 等[12]提出将目标端内领域的单语语料复制一份作为源端数据,与其构成伪平行数据,将得到的伪平行数据与外领域的平行语料进行连接,共同训练一个领域适应的 Transformer 系统,并且证明了通过复制目标端文本到源端得到的伪平行语料不会消减外领域数据训练的模型性能.

d) 基于领域感知特征嵌入的无监督领域适应(domain aware feature embedding, DAFE) Dou 等[15]通过将特定领域的特征嵌入添加到 NMT 编码端的每一层网络中,并且结合语言模型进行多任务学习来训练特定的领域特征.这种在多任务学习网络框架中的训练模型,既有领域外的平行语料,也有通过反向翻译生成的内领域的伪平行语料.本文提出的文化风格区分的方法不仅在编码端添加源语言端的文化风格区分标记,同时在解码端添加目标端语言的文化风格区分标记.本文的文化风格区分标记的特征参数不是通过单独的语言模型进行训练,而是与网络中其他参数一起训练,降低了网络训练的复杂度,提升了模型在电商领域数据的翻译性能.

3 实 验

3.1 实验数据集

从语言数据联盟(LDC)中抽取新闻领域的中英平行语句对训练中英和英中基准系统,训练数据包含 125 万平行语句对,该训练数据集为外领域平行语料.测试集为美国国家标准与技术研究院 2002 年的数据 NIST02、NIST03、NIST04、NIST05 和 NIST08,共 5 个测试数据集.验证集为 NIST06.中文词表大小为 4 万,英文词表大小为 5 万,其余低频词用<UNK>替换.

电商数据集中,因为淘宝和亚马逊网站的电商产品资源非常丰富,并且获取的数据比较具有权威性,所以中文电商领域的产品信息语料取自淘宝官方网站,英文电商领域的产品信息语料取自亚马逊官方网站.中英文数据均主要包括女士服装、男士服装、玩具和食物四大类别,具体的数据统计信息如表 1 所示.

表 1 电商领域的数据统计

Tab. 1 Data statistics of e-commerce

数据集	翻译方向	女士服装	男士服装	玩具	食物	总和
训练集	中文	7.0×10^5	9.0×10^5	3.1×10^5	5.9×10^5	2.5×10^6
	英文	6.2×10^5	6.7×10^5	6.4×10^5	7.0×10^5	2.6×10^6
验证集	中文	587	731	429	780	2 527
	英文	591	588	491	462	2 132
测试集	中文	478	502	468	661	2 109
	英文	417	547	450	580	1 994

注:表中数字为数据集的句子数量.

3.2 实验参数

本文中所有实验均基于开源代码 Fairseq[23],将模型设置为 Transformer,模型的失活率设置为 0.3,编码器和解码器层数均为 6 层,其他基本的超参数设置为 Fairseq 中的默认参数选项,最大保存模型数目设置为 5.解码时,采用集束搜索,其中束大小设置为 5,其余参数采用默认设置.训练和测试均在 NVIDIA TITAN XP GPU 硬件上实现.

3.3 评测标准

双语互译评估[24](bilingual evaluation understudy, BLEU)是一种 MT 的自动评估指标,用来评估 MT 的译文质量,计算公式为:

$$\mathrm{BLEU}=V_{\mathrm{BP}}\times\exp\Big(\sum_{n=1}^{N}w_n\log(p_n)\Big). \tag{3}$$

其中:V_{BP}表示过短惩罚系数,当译文的句子过短时,会给其一定的惩罚;p_n 为 n 元语法的精度,表示译文句子的词出现在参考答案中的概率;w_n 为每个 p_n 的权重.

3.4　不同字节对编码(BPE)实验

由于电商数据多为基于名词实体或短语的堆叠,比如:品牌名和产品的形状等,固定的词表大小产生的未登录词较多. 在将数据用于相关实验之前,本文中使用 BPE[25]技术处理了所有数据. 并且分别设置了不同的 BPE 进行实验,以探索合适的 BPE. 实验结果如表 2 所示,当编码方式为中英单独编码、BPE 为 64 000时,翻译性能最佳,故以下实验均采用中英单独编码,BPE 大小选为 64 000.

3.5　混合训练实验

本文中分别尝试不同比例的电商内领域的伪数据和外领域的平行语料进行实验,得到的实验结果如表 3 所示. 对比添加不同比例的外领域平行语料时 Tranformer 的翻译性能可知,当电商数据与外领域平行语料的数据比例为 1∶1 时,翻译效果最好,这与 Sennrich 等[11]得到的结论一致.

表 2　不同 BPE 的实验结果对比

Tab. 2　Experimental results comparison of different BPE

编码方式	BPE/10^3	翻译方向	BLEU/%				
			女士服装	男士服装	玩具	食物	平均分数
单独编码	64	英中	10.77	11.16	13.30	15.76	12.75
		中英	7.00	7.34	12.02	12.66	9.76
单独编码	32	英中	10.55	10.64	12.43	15.74	12.34
		中英	6.04	6.68	11.92	13.11	9.44
联合编码	64	英中	10.44	11.11	10.47	13.40	11.36
		中英	6.78	6.89	11.26	13.17	9.53

表 3　不同比例数据的实验结果

Tab. 3　Experimental results of different scale data

训练方法	翻译方向	BLEU/%				
		女士服装	男士服装	玩具	食物	平均分数
1∶3	英中	10.18	12.11	13.79	15.96	13.01
	中英	9.43	11.17	16.72	13.73	12.76
1∶2	英中	10.61	11.52	13.88	16.15	13.04
	中英	9.76	12.01	19.45	17.86	14.77
1∶1	英中	11.28	12.87	14.76	17.49	14.10
	中英	10.87	11.94	19.27	18.92	15.25
2∶1	英中	11.12	11.34	13.70	17.07	13.31
	中英	10.85	11.35	13.39	19.00	13.65
混合训练方法	英中	13.21	15.52	17.06	21.17	16.74
	中英	14.34	12.29	19.25	21.29	16.79

注:表中比值为电商数据规模与 LDC 数据规模的比值.

采用本文提出的混合训练方法得到的实验结果记录在表 3 中最后两行. 与以 1∶1 的比例添加外领域平行数据的翻译性能相比,本文提出的混合训练方法仅用电商领域的单语数据及其解码得到的伪数据在英中和中英翻译中平均 BLEU 值分别提升 2.64 和 1.54 个百分点. 虽然未使用质量较高的外领域的

平行语料,本文的混合训练方法相比于其他已有的方法依旧得到了较大的提升. 这与 Edunov 等[21]和 Bogoychev 等[22]的研究结果一致,本文提出的混合训练方法结合了两者的思想,通过混合训练的方法使得模型能够学习到亚马逊和淘宝电商平台产品数据的共同特点,特别是对于同类别产品的数据信息,使得模型能够捕获到不同平台数据的相关信息,从而进一步提升了电商领域的产品信息的翻译效果.

3.6 混合训练+文化风格区分实验

基于混合训练方法添加文化风格特征嵌入的实验结果和相关方法基准系统的实验结果如表 4 所示.

表 4 中,基准系统为仅使用外领域平行语料训练得到的 Transformer 模型. 对比已有的主流方法反向翻译、基于外领域模型微调的反向翻译、目标端复制和 DAFE 方法的翻译性能,可以看出以上实验方法均能有效地提升电商领域产品信息翻译的效果. 其中,对于英中翻译方向,目标端复制方法相对于基准系统的平均 BLEU 值提升最高,为 3.63 个百分点;对于中英翻译方向,基于外领域模型微调的反向翻译方法的性能提升最明显,平均 BLEU 值提升 6.90 个百分点. 本文提出的混合训练方法在英中翻译方向上 BLEU 平均得分为 16.74%,高出基准系统 3.99 个百分点,中英翻译方向上平均 BLEU 得分为 16.79%,高出基准系统 7.03 个百分点,同时相较于已经存在的相关主流方法均有进一步的提升.

表 4 不同方法的实验结果

Tab. 4 Experimental results of different methods

系统	翻译方向	BLEU/%				
		女士服装	男士服装	玩具	食物	平均分数
基准系统	英中	10.77	11.16	13.30	15.76	12.75
	中英	7.00	7.34	12.02	12.66	9.76
反向翻译	英中	11.28	12.87	14.76	17.49	14.10
	中英	10.87	11.94	21.27	18.92	15.75
基于外领域模型微调的反向翻译	英中	12.45	14.73	16.59	19.93	15.93
	中英	13.52	12.50	18.89	21.72	16.66
目标端复制	英中	15.87	13.63	18.19	17.78	16.38
	中英	16.02	12.48	19.42	18.27	16.55
DAFE	英中	13.83	13.69	16.18	18.73	15.61
	中英	13.37	12.47	16.65	19.57	15.52
混合训练	英中	13.21	15.52	17.06	21.17	16.74
	中英	14.34	12.29	19.25	21.29	16.79
混合训练+文化风格特征区分方法	英中	13.07	15.66	18.15	21.41	17.07
	中英	14.28	13.67	19.92	21.46	17.33

从表 4 中最后 2 行的结果来看,在混合训练的基础上增加文化风格特征区分后,在英中翻译方向上的平均 BLEU 得分为 17.07%,高出反向翻译方法 2.97 个百分点,高出目标端复制单语数据的方法 0.69 个百分点,并且相较于混合训练方法,模型效果有进一步地提升;在中英翻译方向上,混合训练+文化风格特征区分在四大类测试集数据上的平均 BLEU 得分为17.33%,高出反向翻译方法 1.58 个百分点,高出目标端复制单语数据的方法 0.78 个百分点. 实验结果表明,文化风格特征嵌入和混合训练的方法相结合使得模型不仅能够学习到同一产品的相关描述,同时也能够捕获到同一产品的不同风格特征的描述,使得模型的翻译效果得到进一步的提升.

4 结 论

针对电商产品翻译系统的训练存在两个主要的问题:电商领域训练数据稀缺和电商产品描述文化风格差异较大,本文将获取的大量产品信息预处理后作为训练语料,并且提出了基于无监督领域适应的混合

训练添加文化风格特征区分的方法.实验结果表明,本文提出的方法提高了基于单语语料的电商产品翻译的准确度.目前仅对于中文淘宝官方网站的电商产品和英文亚马逊官方网站的电商产品数据信息进行了相关实验,未来工作中,将获取更多平台和更多语种的电商领域产品数据信息进行相关研究,使得电商产品信息翻译模型性能够得到进一步的提升.

参考文献:

[1] 李亚超,熊德意,张民. 神经机器翻译综述[J]. 计算机学报,2018,41(12):100-121.

[2] 高明虎,于志强. 神经机器翻译综述[J]. 云南民族大学学报(自然科学版),2019,28(1):76-80.

[3] 刘洋. 神经机器翻译前沿进展[J]. 计算机研究与发展,2017,54(6):1144-1149.

[4] GEHRING J,AULI M,GRANGIER D,et al. Convolutional sequence to sequence learning[C]// International Conference on Machine Learning. Sydney:PMLR,2017:1243-1252.

[5] WU Y,SCHUSTER M,CHEN Z,et al. Google's neural machine translation system: bridging the gap between human and machine translation[EB/OL]. [2020-05-04]. http://arxiv.org/abs/1609.08144.

[6] HASSAN H,AUE A,CHEN C,et al. Achieving human parity on automatic chineseto english news translation [EB/OL]. [2020-05-04]. https://arxiv.org/abs/1803.05567v2.

[7] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]// Advances in Neural Information Processing Systems. Long Beach: NIPS, 2017:5998-6008.

[8] CHU C, DABRE R, KUROHASHI S. An empirical comparison of domain adaptation methods for neural machine translation[C]// Annual Meeting of the Association for Computational Linguistics. Vancouver: ACL,2017:385-391.

[9] VILAR D. Learning hidden unit contribution for adapting neural machine translation models[C]// Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Louisiana:ACL,2018:500-505.

[10] FREITAG M,ALONAIZAN Y. Fast domain adaptation for neural machine translation[EB/OL]. [2020-05-04]. http://arxiv.org/abs/1612.06897.

[11] SENNRICH R, HADDOW B, BIRCH A. Improving neural machine translation models with monolingual data [C]// Annual Meeting of the Association for Computational Linguistics. Berlin:ACL,2016:86-96.

[12] CURREY A,BARONE A V M,HEAFIELD K. Copied monolingual data improves low-resource neural machine translation[C]// Conference on Machine Translation. Copenhagen:ACL,2017:148-156.

[13] ZENG J L,LIU Y,SUN J S,et al. Iterative dual domain adaptation for neural machine translation [C]// Conference on Empirical Methods in Natural Language Processing/International Joint Conference on Natural Language Processing. Hong Kong:ACL,2019:845-855.

[14] CHU C H,RAJ D. Multilingual multi-domain adaptation approaches for neural machine translation. [EB/OL]. [2020-05-04]. https://arxiv.org/abs/1906.07978.

[15] DOU Z Y,HU J J,ANTONIOS A,et al. Unsupervised domain adaptation for neural machine translation with domain-aware feature embeddings[C]// Conference on Empirical Methods in Natural Language Processing/ International Joint Conference on Natural Language Processing. Hong Kong:ACL,2019:1417-1422.

[16] YANG Z, CHEN W, WANG F, et al. Unsupervised domain adaptation for neural machine translation[C]// International Conference on Pattern Recognition. Beijing:IEEE,2018:338-343.

[17] SU J S,ZENG J L,XIE J,et al. Exploring discriminative word-level domain contexts for multi-domain neural machine translation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 43(5): 1530-1545.

[18] ZENG J L. Multi-domain neural machine translation with word-level domain context discrimination[C]// Conference on Empirical Methods in Natural Language Processing. Brussels:ACL,2018. 447-457.

[19] SHOETSU S, JIN S, NAOKI Y, et al. Vocabulary adaptation for distant domain adaptation in neural machine translation[EB/OL]. [2020-05-04]. https://arxiv.org/abs/2004.14821.

[20] GORDONM A,DUHK,DUH K. Distill,adapt,distill: training small, in-domain models for neural machine translation[C]// Annual Meeting of the Association for Computational Linguistics. Vancouver:ACL,2020:110-118.

[21] EDUNOV S, OTT M, RANZATO M, et al. On the evaluation of machine translation systems trained with back-translation[EB/OL]. [2020-05-04]. http://arxiv.org/abs/1908.05204.

[22] BOGOYCHEV N, SENNRICH R. Domain, translationese and noise in synthetic data for neural machine translation [EB/OL]. [2020-05-04]. https://arxiv.org/abs/1911.03362.

[23] MYLE O,SERGEY E,ALEXEI B,et al. Fairseq:a fast,

extensible toolkit for sequence modeling [C] // Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis: ACL, 2019: 48-53.

[24] KISHORE P, SALIM R, TODD W, et al. BLEU: a method for automatic evaluation of machine translation [C] // Annual Meeting of the Association for Computational Linguistics. Philadelphia: ACL, 2002: 311-318.

[25] SENNRICH R, HADDOW B, BIRCH A. Neural machine translation of rare words with subword units [C] // Annual Meeting of the Association for Computational Linguistics. Berlin: ACL, 2016: 1715-1725.

Culture-style aware e-commerce product translation based on unsupervised domain adaptation

SHI Xiaojing, NING Qiuyi, DUAN Xiangyu*

(School of Computer Science and Technology, Soochow University, Suzhou 215006, China)

Abstract: Generally, two major problems in the training of e-commerce product translation system are encountered, namely, the scarcity of training data in the e-commerce field and the difference in cultural-style of e-commerce product description. In order to improve the performance of e-commerce product translation system, we have collected a large amount of product data information as training corpus and propose mix-training and culture-style aware methods based on unsupervised domain adaptation. In the mix-training method, we mix the pseudo corpus of the e-commerce domain data obtained by the model system based on external domain data training to obtain a new model. In the method of cultural-style aware, we add corresponding cultural-style distinction marks to e-commerce data of different languages, tell the model of current data category in the training process, and obtain the corresponding cultural style distinguishing feature vector according to the category information. Experimental results indicate that our method outperforms various existing e-commerce product translations based on monolingual corpus.

Keywords: machine translation; domain adaptation; unsupervised