基于短语替换的汉越伪平行句对生成

贾承勋, 赖华, 余正涛, 文永华, 于志强
1. 昆明理工大学 信息工程与自动化学院；2. 昆明理工大学 云南省人工智能重点实验室

论文摘要

神经机器翻译在语料丰富的语种上取得了良好的翻译效果，但是在汉语-越南语这类双语资源稀缺的语种上性能不佳，通过对现有小规模双语数据进行词级替换生成伪平行句对可以较好地缓解此类问题。考虑到汉越词级替换中易存在一词多译问题，所以对基于更大粒度的替换进行了研究，提出了一种基于短语替换的汉越伪平行句对生成方法。利用小规模双语数据进行短语抽取构建短语对齐表，并通过在维基百科中抽取的实体词组对其进行扩充，在对双语数据的汉语和越南语分别进行短语识别后，利用短语对齐表中与识别出的短语相似性较高的短语对进行替换，以此实现短语级的数据增强，并将生成的伪平行句对与原始数据一起训练最终的神经机器翻译模型。在汉-越翻译任务上的实验结果表明，通过短语替换生成的伪平行句对可以有效提高汉-越神经机器翻译的性能。

论文简介

神经机器翻译(NMT)是近几年提出的机器翻译方法，在大量的平行句对中取得了很好的翻译效果，并超越了统计机器翻译(SMT)，但对于资源稀缺型语言的效果并不理想。

利用现有小规模数据生成伪平行数据是提升资源稀缺语言神经机器翻译最有效的途径之一。目前扩充伪平行数据的方法主要有四种，第一种是抽取式方法，在大量的可比语料中抽取平行句对，将源语言句子和目标语言句子放到同一语言空间下，通过比较源语言句子和目标语言句子之间的相似性判断句对是否平行；第二种是目前应用较为广泛的回译方法(Back Translation, BT)，使用目标语言单语数据通过目标到源语言的翻译模型翻译为源语言译文，从而生成伪平行数据；第三种方法是利用枢轴语言的方法，通过训练源语言到枢轴语言模型和枢轴语言到目标语言模型，利用源语言单语数据通过两步翻译生成伪平行数据进行源语言到目标语言的模型训练；第四种方法是在现有双语数据基础上进行词的替换，通过对现有数据进行分析，在一定的替换规则下进行词或模块间的替换，以此生成更多的伪平行句对达到数据增强的效果，但**词级替换易出现一词多译的问题，从而对生成的伪平行句对产生不良的影响**。

目前词级替换的数据增强中存在一词多义的问题，考虑到短语产生歧义的可能较小，并且可以包含更多的可利用信息，可以缓解一词多译带来的不良影响，同时进行短语替换可以使生成的伪平行句对包含更完整的结构和信息，因此本文对短语进行识别并根据一定的替换规则实现短语替换。

通过对原始小规模数据进行短语抽取构建出一对一翻译的短语对齐表，由于仅利用原始数据抽取的短语不足以生成增加数据的多样性，因此我们对短语对齐表补入了基于维基百科抽取出的汉越实体词组，在对数据中的短语进行识别后，通过计算余弦相似度计算出与替换短语相似的短语，最后通过短语对齐表对替换位置的短语进行替换，将生成的伪平行句对与原始数据一起训练最终汉-越神经机器翻译模型。

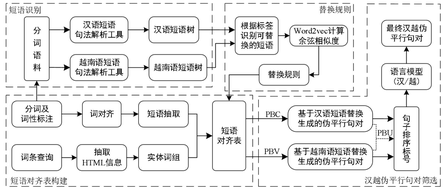
系统模型

实验所用的基准模型是谷歌(Google)开源的神经机器翻译模型Transformer。

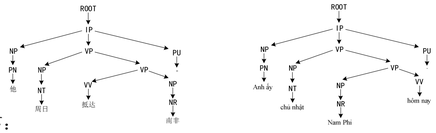
词表大小设置为30K，实验均在单卡GPU服务器上进行，为防止出现过拟合现象，在多次试验调整后将dropout值设置为0.1，batch size为64，hidden units为512，train steps为200K。

使用BLEU4作为评测指标。

方法原理



首先对原始数据进行短语抽取构建短语对齐表，并用维基百科词条中抽取的人名、地名、组织机构名和专有名词等实体词组对短语对齐表进行扩充，然后利用分词后的数据通过短语句法解析工具进行可替换短语的识别，其次我们还在抽取命名实体的时候，会将其相应的摘要抽取出来分为汉语和越南语存储为文本数据，与原始的汉语和越南语数据分别映射到向量空间中，计算出向量空间中可替换短语的相似短语进行替换，同时考虑到汉语和越南语在句法上存在差异性，因此我们分别进行以汉语短语为基础的替换(Chinese phrase based substitute, PBS-C)和以越南语为基础的替换(Vietnamese phrase based substitute, PBS-V)。

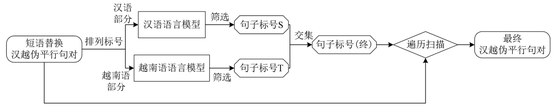


短语替换规则总结如下：

(1)对一个句子只进行一次替换；

(2)将N值设置为5，即只选择利用相似性前5的短语进行替换；

(3)只对相同词性的短语进行替换，即名词短语(NP)只利用名词短语替换；



利用语言模型筛选数据的特点在于首先对生成的伪平行句对进行排序标号，然后分别利用汉语语言模型和越南语语言模型对伪平行句对各自语言部分进行困惑度评判，过滤掉困惑度大于阈值的句子序号，然后根据最终保留下来的句对序号，将汉语和越南语部分的序号取交集，最后在原始的伪平行数据中遍历扫描，得到最终可用的伪平行句对。

实验仿真

为了验证短语替换与词级替换生成的伪平行句对的性能提升差异，对同义词替换生成伪平行句对的方法(EDA)进行了对比。其中汉-越语料通过网络爬取收集并清洗后得到154K平行句对，最终保留长度在50个词以内的句对，从中分别随机抽取2000平行句对作为测试集和验证集。实验数据的准备情况如表1所示。

表1 实验数据准备情况

	训练集	测试集	验证集
汉-越	150K	2K	2K

实验主要研究了通过短语替换生成伪平行句对对汉越神经机器翻译性能的影响，baseline为仅利用原始数据训练得到的模型翻译效果，对比了通过汉语进行短语替换生成伪平行句对、通过越南语生成伪平行句对以及对通过汉语和越南语生成的伪平行句对取并集的方式对系统性能提升的影响，经过语言模型筛选后，PBS-C(基于汉语短语替换)和PBS-V(基于越南语短语替换)最终分别留下了约450K的伪平行句对，PBS-U(短语替换取结果并集)留下了约670K的伪平行句对，结果如表2所示。

表2 添加伪平行句对后的实验结果

method	伪平行句对	总句对规模	汉-越	越-汉
baseline	—	150K	21.19	22.03
PBS-C	450K	150K+450K	22.38	23.54
PBS-V	450K	150K+450K	22.61	23.26
PBS-U	670K	150K+670K	**22.78**	**23.75**

通过以上实验可以看出，通过短语替换生成的伪平行句对可以提升神经机器翻译模型的翻译性能。为了验证基于短语替换生成的伪平行句对相较于基于词级的替换可以更好地提升系统的翻译性能，在此通过同义词替换生成伪平行句对，为了更好的控制影响因素，统一通过汉语进行替换，并且将生成的伪平行句对量统一为450K，实验结果如表3所示。

表3 添加伪平行句对后的实验结果

method	baseline	EDA-同义词替换	PBS-C
汉-越	21.19	22.01	**22.38**
越-汉	22.03	22.94	**23.54**

由实验结果可以看出，同义词替换生成的伪平行句对比通过短语替换生成的伪平行句对对系统性能的提升较低，这是因为基于短语替换解决了一词多译问题，生成的伪平行句对包含更多的可用有效信息，进一步提升了汉越神经机器翻译模型的性能。

论文结论

针对汉越神经机器翻译的数据稀缺问题，考虑到短语不易存在歧义且包含更多可利用信息，因此提出了通过短语替换进行汉越伪平行句对生成的方法。利用在原始数据中抽取出的短语和维基百科中抽取的实体词组构建短语对齐表，根据余弦相似性在短语对齐表中选择与可替换短语相近的短语进行替换，以此生成新的汉越伪平行句对，提升汉越神经机器翻译的性能。实验结果表明，这种方法可以在汉越神经机器翻译中更好的提升模型的翻译性能。在未来工作中，我们会探索对可替换短语的扩充及短语识别的准确性对汉-越神经机器翻译性能的影响。

西安电子科技大学 XIDIAN UNIVERSITY