(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 112926345 A
(43)申请公布日 2021.06.08

(21)申请号 202110378632.1

(22)申请日 2021.04.08

(71)申请人 中国科学技术大学
地址 230026 安徽省合肥市包河区金寨路96号

(72)发明人 陈贝多 黄青青 杜俊

(74)专利代理机构 北京凯特来知识产权代理有限公司 11260
代理人 郑立明 韩珂

(51)Int.Cl.
G06F 40/58(2020.01)
G06K 9/62(2006.01)

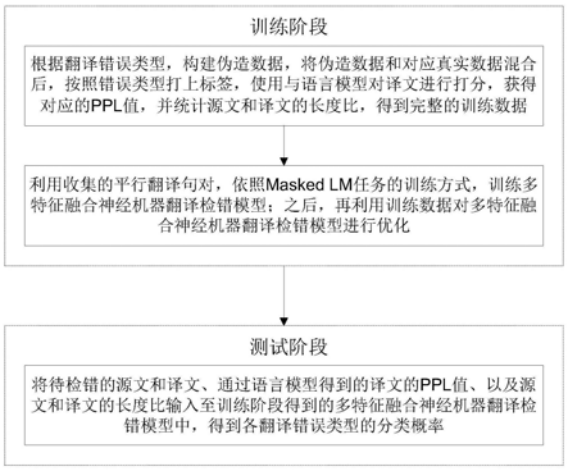
权利要求书1页 说明书8页 附图1页

(54)发明名称

基于数据增强训练的多特征融合神经机器翻译检错方法

(57)摘要

本发明公开了一种基于数据增强训练的多特征融合神经机器翻译检错方法，针对真实错误句对进行人工分类和归纳，通过数据增强的方法，伪造大量数据，增强机器翻译检错模型效果和鲁棒性，并且在模型输入中加入源文译文长度比值信息、译文语言模型PPL分数特征信息，从而进一步提升检错模型分类准确率，基于该检错方案，检测结果可以用于后续纠错，也可用于错误提示，提供翻译用户体验；还可以用于机器翻译效果的评价指标。



CN 112926345 A

1.一种基于数据增强训练的多特征融合神经机器翻译检错方法,其特征在于,包括:

训练阶段:根据翻译错误类型,构建伪造数据,将伪造数据和对应真实数据混合后,按照错误类型打上标签,使用与语言模型对译文进行打分,获得对应的PPL值,并统计源文和译文的长度比,得到完整的训练数据;利用收集的平行翻译句对,依照Masked LM任务的训练方式,训练多特征融合神经机器翻译检错模型;之后,再利用训练数据对多特征融合神经机器翻译检错模型进行优化;

测试阶段:将待检错的源文和译文、通过语言模型得到的译文的PPL值、以及源文和译文的长度比输入至训练阶段得到的多特征融合神经机器翻译检错模型中,得到各翻译错误类型的分类概率。

2.根据权利要求1所述的一种基于数据增强训练的多特征融合神经机器翻译检错方法,其特征在于,通过人工进行数据收集和标注,划分翻译错误类型;针对每种翻译错误类型,基于规则或相关算法,使用计算机技术进行数据伪造,从而构建相应的伪造数据。

3.根据权利要求1或2所述的一种基于数据增强训练的多特征融合神经机器翻译检错方法,其特征在于,所述根据翻译错误类型,构建伪造数据包括:

基于平行翻译句对,使用基于统计方法的词对齐工具预先完成句对中的词对齐,基于命名实体识别工具进行人名、地名和数字的标注,以及使用句法分析模型构建句子依存句法结构信息;

然后,针对每类翻译错误类型,基于词对齐信息、命名实体识别信息以及句法结构信息,使用规则对平行翻译句对进行修改,并结合机器翻译构建伪造的机器翻译错误类型数据。

4.根据权利要求1或2所述的一种基于数据增强训练的多特征融合神经机器翻译检错方法,其特征在于,所述翻译错误类型包括如下类别中的一种或多种:译文多译、译文漏译、直译错误、重复翻译、标点符号错误、英文大小写错误、中文拼音翻译错误。

5.根据权利要求1所述的一种基于数据增强训练的多特征融合神经机器翻译检错方法,其特征在于,训练数据的数据格式为:源文-译文-译文PPL值-长度比-翻译错误类型标签。

6.根据权利要求1所述的一种基于数据增强训练的多特征融合神经机器翻译检错方法,其特征在于,所述利用收集的平行翻译句对,依照Masked LM任务的训练方式,训练多特征融合神经机器翻译检错模型包括:

对于平行翻译句对,随机遮蔽连续片段的源文或者译文,从而引导多特征融合神经机器翻译检错模型学习平行翻译句对之间的翻译对应关系。

7.根据权利要求1所述的一种基于数据增强训练的多特征融合神经机器翻译检错方法,其特征在于,训练阶段与测试阶段中,将译文的PPL值、以及源文和译文的长度比量化后,再输入至多特征融合神经机器翻译检错模型;

量化方式为:统计PPL值和长度比值分布,各自划分为多个离散区间,按照离散区间将当前的译文的PPL值、以及源文和译文的长度比量化进对应的离散区间。

基于数据增强训练的多特征融合神经机器翻译检错方法

技术领域

[0001] 本发明涉及机器翻译检错技术领域，尤其涉及一种基于数据增强训练的多特征融合神经机器翻译检错方法。

背景技术

[0002] 机器翻译，又称为自动翻译，是指利用计算机将一种自然语言(源语言)转换为另一种自然语言(目标语言)的过程。机器翻译具有重要的实用价值。

[0003] 目前神经网络框架的机器翻译效果得到显著提升，例如:1)Devlin J,Chang M W,Lee K,et al.BERT:Pre-training of Deep Bidirectional Transformers for Language Understanding[J].2018;2)Bahdanau D,Cho K,Bengio Y.Neural Machine Translation by Jointly Learning to Align and Translate[J].Computer ence,2014;3)Jonas Gehring,Michael Auli,David Grangier,Denis Yarats,and Yann N.Dauphin.Convolutional sequence to sequence learning.arXiv preprint arXiv:1705.03122v2,2017。

[0004] 基于注意力机制的Transformer框架(A.Vaswani,N.Shazeer,N.Parmar,J.Uszkoreit,L.Jones,A.N.Gomez,L.Kaiser,and I.Polosukhin,“Attention is all you need,”CoRR,vol.abs/1706.03762,2017.)在翻译任务上大显身手，目前已成为工业界主流机器翻译技术，在通用场景下机器翻译基本可用，但是基于数据驱动的神经机器翻译技术仍然存在许多问题，如口语化翻译错误、专业领域翻译错误以及低资源语种翻译问题，当前技术条件下还没有达到理想水平，机器翻译系统仍然会输出的错误的翻译结果。

[0005] 机器翻译结果检错，是指对给定的句子判断其是否存在翻译错误，以及错误类型。目前主要采用如下几种方案：

[0006] 1、人工检错方法

[0007] 人工根据所学知识及经验进行错误检测、标注错误信息。人工检错示例如表1所示。

[0008]

问题点	源文	译文	正确译文
1.音译错误	Scrubby.	斯克鲁比。	灌木丛生的。
2.专有名词/人名直译	Um, when I operate the radio, it goes to DVD. But now I don't know how to put the reverse camera back on.	嗯，当我操作收音机时，它会转到DVD。但现在我不知道怎么把倒车摄像头装回去。	嗯，当我收听收音机时，它会转到DVD。但现在我不知道怎么把倒车影像装回去。
	They decided to book the air b and b.	他们决定预订空中食宿。	他们决定预订民宿。
3.多译漏译	肚子疼。	I have a stomachache.	Stomachache.
	热烈鼓掌。	Applaud.	Applaud warmly.

[0009] 表1人工检错示例

[0010] 2、基于规则的机器翻译结果检错方法。

[0011] 该方法基于人工设计规则，或则从数据中抽取规则实现机器翻译的自动检错，适用于特定任务和领域问题进行检错，主要有以下几种方法：

[0012] 1)模板匹配法。

[0013] 在机器翻译中针对特定应用场景(例如购票)，翻译结果可能比较受限，对于常出现的错误人工总结一些错误模板，用于对翻译结果的检错处理。例如“我想购买从合肥到上海的机票”中“从*到*的机票”就可以认为是一个模板，如果检测出“从*道*的机票”则可以认为出现错误。此外这些错误的模板也可以根据机器翻译结果数据和标注答案数据的对比，自动统计得到。在某些受限领域该方法可以作为一种常用的检错方式。

[0014] 2)基于上下文语义分析

[0015] 该方法主要是根据语法规则和上下文约束实现检错功能。例如“我国今后每年都要今后许多小麦”中第二个“今后”明显不符合语法规则，可以认为是错误。规则和约束关系通常都需要人工来编写，实际应用中主要针对特定一类错误进行检错，例如“数字+量词+名词”的规则类。

[0016] 3、基于统计的翻译结果检错方法。

[0017] 由于基于规则的检错方案通常只能对于特定任务和领域问题进行检错。其他情况可以使用统计的方法用于机器翻译自动检错处理。主要有以下几种：

[0018] 1)基于语言模型的检错方法。

[0019] 该方法采用语言模型来统计每个词上下文的语境信息，如果某句话或者某个词的语言模型信息很大，那么则认为该句存在错误，语言模型得分很大的词可能存在错误。

[0020] 2)基于词语共现的检错方法。

[0021] 该方法基于大量的语料库，设定一个窗口长度W，定义在中心词周围W/2的词为其邻居。通过遍历文本统计每个词的邻居和出现次数，并对邻居进行排序，得到每个词的邻居特征。对于翻译结果中的每个可能错误词，计算当前句子邻域特征与基于语料库的统计邻域特征之KL距离作为该词置信度得分，然后通过词性相似寻找该词的混淆词候选，同样计算候选混淆词在当前句子中的置信度得分，最后得出当前词置信度与潜在的候选混淆词的置信度得分，确定是否发生错误并给出正确词。

[0022] 但是，现有的上述翻译检错方案存在着效率低下、成本高昂、人力浪费等问题，已逐渐无法满足快速增长的翻译需求。

发明内容

[0023] 本发明的目的是提供一种基于数据增强训练的多特征融合神经机器翻译检错方法，具有较高的灵活性，并且节省人力以及时间，能够快速的发现翻译结果错误并及时纠正或提示，提高机器翻译的准确度和交互体验。

[0024] 本发明的目的是通过以下技术方案实现的：

[0025] 一种基于数据增强训练的多特征融合神经机器翻译检错方法，包括：

[0026] 训练阶段：根据翻译错误类型，构建伪造数据，将伪造数据和对应真实数据混合后，按照错误类型打上标签，使用与语言模型对译文进行打分，获得对应的PPL值，并统计源文和译文的长度比，得到完整的训练数据；利用收集的平行翻译句对，依照Masked LM任务的训练方式，训练多特征融合神经机器翻译检错模型；之后，再利用训练数据对多特征融合

神经机器翻译检错模型进行优化；

[0027] 测试阶段：将待检错的源文和译文、通过语言模型得到的译文的PPL值、以及源文和译文的长度比输入至训练阶段得到的多特征融合神经机器翻译检错模型中，得到各翻译错误类型的分类概率。

[0028] 由上述本发明提供的技术方案可以看出，针对真实错误句对进行人工分类和归纳，通过数据增强的方法，伪造大量数据，增强机器翻译检错模型效果和鲁棒性，并且在模型输入中加入源文译文长度比值信息、译文语言模型PPL分数特征信息，从而进一步提升检错模型分类准确率，基于该检错方案，检测结果可以用于后续纠错，也可用于错误提示，提供翻译用户体验；还可以用于机器翻译效果的评价指标。

附图说明

[0029] 为了更清楚地说明本发明实施例的技术方案，下面将对实施例描述中所需要使用的附图作简单地介绍，显而易见地，下面描述中的附图仅仅是本发明的一些实施例，对于本领域的普通技术人员来讲，在不付出创造性劳动的前提下，还可以根据这些附图获得其他附图。

[0030] 图1为本发明实施例提供的一种基于数据增强训练的多特征融合神经机器翻译检错方法的流程图；

[0031] 图2为本发明实施例提供的多特征融合的神经机器翻译纠错模型框架示意图。

具体实施方式

[0032] 下面结合本发明实施例中的附图，对本发明实施例中的技术方案进行清楚、完整地描述，显然，所描述的实施例仅仅是本发明一部分实施例，而不是全部的实施例。基于本发明的实施例，本领域普通技术人员在没有做出创造性劳动前提下所获得的所有其他实施例，都属于本发明的保护范围。

[0033] 在机器翻译应用中，现有的翻译结果检错方案利用人工检错是一种可行的途径，但是现有方案过于依赖人工干预，需要先总结错误类型以及收集大量语料库，效率低下。为此，本发明实施例提供一种基于数据增强训练的多特征融合神经机器翻译检错方法，首先对现有的错误翻译句对数据进行人工归类和总结特征，基于翻译错误类型，提出一种数据增强方案来优化检错模型，针对每类翻译错误类型构建大量的伪造数据，将伪造数据和真实数据按照错误类型打上标签，并使用语言模型对译文进行打分，统计源文和译文长度比，从而获取译文PPL值和长度比，基于上述过程完成训练数据的构建。此后，将译文PPL值和长度比作为辅助特征进行翻译检错模型的训练，完成训练后即可进行机器翻译结果的错误类型判断。

[0034] 本发明实施例上述方案基于翻译句对进行错误类型分类和归纳，并构造大量伪造数据优化检错模型，提升机器翻译检错效果和模型鲁棒性；而翻译检错模型中，额外添加源文译文长度比、译文语言模型PPL值特征信息进行辅助训练，从而进一步提升机器翻译检错准确率。基于上述方案，可以提前发现翻译错误，从而及时反馈翻译结果的错误信息，进行纠正或提示，提高机器翻译系统的准确度和交互体验。

[0035] 如图1所示，本发明实施例提供的上述方案主要包括：

[0036] 一、训练阶段。

[0037] 1、数据增强。

[0038] 根据翻译错误类型，构建伪造数据，将伪造数据和对应真实数据(即机器翻译得到的错误的源文与译文对数据)混合后，按照错误类型打上标签，使用与语言模型对译文进行打分，获得对应的PPL值，并统计源文和译文的长度比，得到完整的训练数据。

[0039] 本发明实施例中，通过人工进行数据收集和标注，划分翻译错误类型；针对每种翻译错误类型，基于规则或相关算法，使用计算机技术进行数据伪造，从而构建相应的伪造数据；具体来说：首先，基于平行翻译句对，使用基于统计方法的词对齐工具(mgiza)预先完成句对中的词对齐，基于命名实体识别工具(NER)进行人名、地名和数字的标注，以及使用句法分析模型构建句子依存句法结构信息；然后，针对每类翻译错误类型，基于词对齐信息、命名实体识别信息以及句法结构信息，使用规则对平行翻译句对进行修改(可以修改源文也可以修改译文)，并结合机器翻译构建伪造的机器翻译错误类型数据。

[0040] 本发明实施例中，所述翻译错误类型包括如下类别中的一种或多种：译文多译、译文漏译、直译错误、重复翻译、标点符号错误、英文大小写错误、中文拼音翻译错误。

[0041] 本发明实施例中，可按照机器翻译训练数据处理流程进行分词、bpe等处理，然后统计每句对应的源文和译文的长度比值。

[0042] 本发明实施例中，训练数据的数据格式为：源文-译文-译文PPL值-长度比-翻译错误类型标签。

[0043] 值得注意的是，语言模型的训练是为了在机器翻译检错模型中，添加语言模型PPL得分特征信息。语言模型结构的选型不是本发明的重点，不失一般性，可以选择基于RNN结构的语言模型进行训练，并能够实现对下一个词预测即可。这里只需要训练目标端的语言模型，具体过程不再赘述。

[0044] 2、模型训练。

[0045] 本发明实施例中，多特征融合神经机器翻译检错模型可以以BERT模型(Jacob Devlin,Ming-Wei Chang,Kenton Lee,and Kristina Toutanova.BERT:Pre-training of deepbidirectional transformers for language understanding.In Proceedings of the 2019Conference ofthe North American Chapter of the Association for Computational Linguistics:Human LanguageTechnologies,Volume 1(Long and Short Papers),pp.4171-4186,Minneapolis,Minnesota,June2019.)为基础框架，输入部分增加机器翻译纠错相关特征，如译文语言模型得分[TM](即译文PPL值)源文和译文长度比信息[L]，另外原BERT模型训练中上下句替换为机器翻译的源文和译文，整体的输入为【译文语言模型得分-TM】【源文译文长度比-L】【源文+[SEP]+译文】，而CLS用于预测对应翻译错误类型的类别。

[0046] 本发明实施例中，模型训练分为两步训练：第一步，利用收集的平行翻译句对，依照Masked LM任务的训练方式，训练多特征融合神经机器翻译检错模型；第二步，用训练数据对多特征融合神经机器翻译检错模型进行优化。下面针对两步训练的过程做相关说明。

[0047] 1)Masked LM任务。

[0048] 在经典BERT模型训练任务中，为了训练深度双向表征，随机遮蔽输入token的某些部分，然后预测被遮住的token。将这一步骤称为「masked LM」(MLM)。在这种情况下，对应遮

蔽token的最终隐藏向量会输入到softmax函数中，并如标准LM中那样预测所有词汇的概率。

[0049] 本发明实施例中，输入为翻译句对且能够让模型学习到源文和目标句翻译对应关系，与BERT训练不同，一次迭代过程中，本发明提出随机mask(遮蔽)掉连续片段的源文或者译文，且由于源文和译文语义存在转换关系，mask比例可以适当提高，从而引导多特征融合神经机器翻译检错模型学习平行翻译句对之间的翻译对应关系。在Masked LM任务训练中，本发明仅使用大量的翻译平行句对进行模型训练。

[0050] 2)翻译错误类型预测任务。

[0051] 在预测翻译错误类型任务中，按照图2模型所示，加入[CLS]、[TM]、[L]进行训练，训练数据按照不同翻译错误类型经验上的分布比例进行数据混合，通过在CLS位置上输出的分类概率计算预测损失。

[0052] 本发明实施例中，额外增加了语言翻译得分信息[TM]和源文译文长度比信息[L]来优化检错模型效果。经验上，语言翻译得分能反映机器翻译译文的流畅度，通过流畅度较低的译文出现错误翻译的概率也越大，而长度比异常的句子常出现严重漏译、多译、重复翻译等问题，因此通过加入以上两种特征信息，能进一步提升检错模型分类准确率。在模型输出部分，在网络输出cls位置上接softmax层计算错误类型概率，输出类别。

[0053] 本发明实施例中，由于PPL值和译文长度比为连续的具体数值，无法作为模型的直接输入，因此，需要将PPL值和长度比进行量化，通过统计机器翻译句对语料库译文PPL值和长度比值分布，各种划分多个(例如，10个)离散区间，并按照离散区间将PPL值和长度比量化进对应区间。

[0054] 示例性的：对于长度比值为0.75量化为0.7，而低于0.6的统一量化为0.6，同理可得PPL值量化。

[0055] 源文译文长度比区间划分：【0.60.7 0.8 0.9 1.0 1.1 1.2 1.3 1.4 1.5】。

[0056] 语言模型PPL得分区间划分：【10 20304050 60 70 80 90 100】。

[0057] 二、测试阶段。

[0058] 将待检错的源文和译文、通过语言模型得到的译文的PPL值、以及源文和译文的长度比输入至训练阶段得到的多特征融合神经机器翻译检错模型中，得到各翻译错误类型的分类概率，如果翻译正确，则各翻译错误类型的分类概率均为0。

[0059] 测试阶段，同样需要对译文的PPL值、以及源文和译文的长度比进行量化，量化方案与训练阶段相同。

[0060] 为了便于前文所述的数据增强方案，下面结合具体的示例对各翻译错误数据的数据增强方案进行说明。

[0061] 构建翻译错误类型1：译文多译。

[0062] 多译错误表现为译文中出现源文中不存在的翻译，本方法基于正确的机器翻译句对和平行翻译句对构建伪造多译错误；

[0063] 人称代词和实体词多译：不改变标准译文，随机去掉源文中他、她、它等人称代词，以及随机删除使用NER标记出来的部分人名、数字和地名。

[0064] 句子成分多译：基于句子依存句法结构信息，随机删除源文中定语、状语等成分。

[0065] 译文句末多译：在前后篇章成句的翻译句对中，对当前句译文后添加下一个句子

译文的子句。

[0066] 如表2所示，为译文多译的数据增强示例。

[0067]

源文	近日，A公司因“二选一”等涉嫌垄断行为，被市场监管总局依法立案调查。此举受到全社会广泛关注，也引发了人们对于平台经济发展的思考。 Recently, A company was investigated by the General Administration of Market Supervision in accordance with the law for suspected monopoly such as "two choices one". This has attracted wide attention from the whole society, and has also aroused people's thinking about the development of platform economy.
构建多译	近日，A公司因“二选一”等涉嫌垄断行为，被市场监管总局依法立案调查。 Recently, A company was investigated by the General Administration of Market Supervision in accordance with the law for suspected monopoly such as "two choices one", this has attracted wide attention.

[0068] 表2译文多译的数据增强示例

[0069] 构建翻译错误类型2：译文漏译。

[0070] 漏译错误表现为译文中漏掉源文中部分存在的翻译，本方法基于正确的机器翻译句对和平行翻译句对构建伪造漏译错误；

[0071] 人称代词和实体词漏译：不改变源文，随机去掉译文中他、她、它等人称代词，以及随机删除使用NER标记出来的部分人名、数字和地名。

[0072] 句子成分漏译：基于句子依存句法结构信息，随机删除译文中定语、状语等成分。

[0073] 译文句末漏译：在前后篇章成句的翻译句对中，对当前句译文后删除上一个句子译文的子句。

[0074] 如表3所示，为译文漏译的数据增强示例。

[0075]

源文	近日，A公司因“二选一”等涉嫌垄断行为，被市场监管总局依法立案调查。 Recently, A company was investigated by the General Administration of Market Supervision in accordance with the law for suspected monopoly such as "two choices one", this has attracted wide attention. 此举受到全社会广泛关注，也引发了人们对于平台经济发展的思考。 This has attracted wide attention from the whole society, and has also aroused people's thinking about the development of platform economy.
构建漏译	近日，A公司因“二选一”等涉嫌垄断行为，被市场监管总局依法立案调查。此举受到全社会广泛关注，也引发了人们对于平台经济发展的思考。 Recently, A company was investigated by the General Administration of Market Supervision in accordance with the law for suspected monopoly such as "two choices one". This has attracted wide attention from the whole society.

[0076] 表3译文漏译的数据增强示例

[0077] 构建翻译错误类型3：直译错误。

[0078] 直译错误常表现为逐词翻译，易出现在长句翻译和新词翻译中。

[0079] 构建新词、短语直译词汇表，将短语、新词对应直译译文在机器翻译系统中进行定制，得到直译译文。

[0080] 将长句源文按照标点或句子成分切分成多个子句，每个子句单独送进机器翻译系统(建议使用SMT机器翻译)进行翻译，然后将译文合并成整句当作译文。

[0081] 如表4所示，为直译错误的数据增强示例。

[0082]

源文	出于呵护新产业、Out of caring for new industries, 新业态的考虑，The consideration of new formats, 我国对平台企业监管一直十分审慎。China has been very cautious in the supervision of platform enterprises.
构建直译	出于呵护新产业、新业态的考虑，我国对平台企业监管一直十分审慎。 Out of caring for new industries, the consideration of new formats, China has been very cautious in the supervision of platform enterprises.

[0083] 表4直译错误的数据增强示例

[0084] 构建翻译错误类型4:重复翻译。

[0085] 重复翻译表现译文中源文重复翻译,本方法基于正确的机器翻译句对和平行翻译句对构建重复翻译。

[0086] 构造常见重复翻译词汇表,例如OK,句子中找到重复翻译词汇在后面添加重复词汇。

[0087] 句子成分重复翻译:基于句子依存句法结构信息,将译文中名词成分重复。

[0088] 源文重复进行机器翻译得到译文,去掉源文重复内容,得到译文重复翻译。

[0089] 如表5所示,为重复翻译的数据增强示例。

[0090]

源文	你好你好你好|Hellohellohello
构建重复翻译	你好|Hellohellohello

[0091] 表5重复翻译的数据增强示例

[0092] 构建翻译错误类型5:中文拼音翻译错误。

[0093] 拼音错误表现为译文中出现人名拼音错误,本方法基于正确的机器翻译句对和平行翻译句对构建伪造多译错误。

[0094] 构造常见人名拼音错误词汇对表,例如Mr.Feng.|Mr.Fung。

[0095] 在句子中找到拼音匹配的词汇替换成后面拼音错误词汇。

[0096] 如表6所示,为中文拼音翻译错误的数据增强示例。

[0097]

源文	冯先生|Mr.Feng.
构建拼音错误	冯先生|Mr.Fung.

[0098] 表6中文拼音翻译错误的数据增强示例

[0099] 构建翻译错误类型6:标点符号错误

[0100] 标点符号错误表现为译文中出现标点符号使用错误,本方法基于正确的机器翻译句对和平行翻译句对构建伪造多译错误。

[0101] 构造常见错误标点对词汇表,例如.|?

[0102] 构造与标点错误词汇联系的正确译文,例如好的。

[0103] 将句子中的末尾标点替换成标点错误词汇表中的标点。

[0104] 如表7所示,为标点符号错误的数据增强示例。

[0105]

源文	好的。|OK.
构建标点错误	好的。|OK?

[0106] 表7标点符号错误的数据增强示例

[0107] 构建翻译错误类型7:大英文大小写错误。

[0108] 大小写错误表现为译文中存在大小写错误的翻译,本方法基于正确的机器翻译句

对和平行翻译句对构建伪造大小写错误。

[0109] 人称代词和实体词大小写错误:构造大小写错误对表,例如:He|heshe|She不改变源文,替换译文中he、she、it等人称代词,以及随机删除使用NER标记出来的部分人名、数字和地名。

[0110] 句子成分大小写错误:基于句子依存句法结构信息,替换译文中名词成分的大小写。

[0111] 缩略词大小写错误:构造大小写错误对表,例如:isn't|ISN't,不改变源文,替换译文中isn't等缩略词。

[0112] 如表8所示,为英文大小写错误的数据增强示例。

[0113]

源文	这件事真的很好笑,对不对呀?|This thing is really funny,isn't it?
构建大小写错误	这件事真的很好笑,对不对呀?|This thing is really funny,ISN't it?

[0114] 表8英文大小写错误的数据增强示例

[0115] 通过以上的实施方式的描述,本领域的技术人员可以清楚地了解到上述实施例可以通过软件实现,也可以借助软件加必要的通用硬件平台的方式来实现。基于这样的理解,上述实施例的技术方案可以以软件产品的形式体现出来,该软件产品可以存储在一个非易失性存储介质(可以是CD-ROM,U盘,移动硬盘等)中,包括若干指令用以使得一台计算机设备(可以是个人计算机,服务器,或者网络设备等)执行本发明各个实施例所述的方法。

[0116] 以上所述,仅为本发明较佳的具体实施方式,但本发明的保护范围并不局限于此,任何熟悉本技术领域的技术人员在本发明披露的技术范围内,可轻易想到的变化或替换,都应涵盖在本发明的保护范围之内。因此,本发明的保护范围应该以权利要求书的保护范围为准。

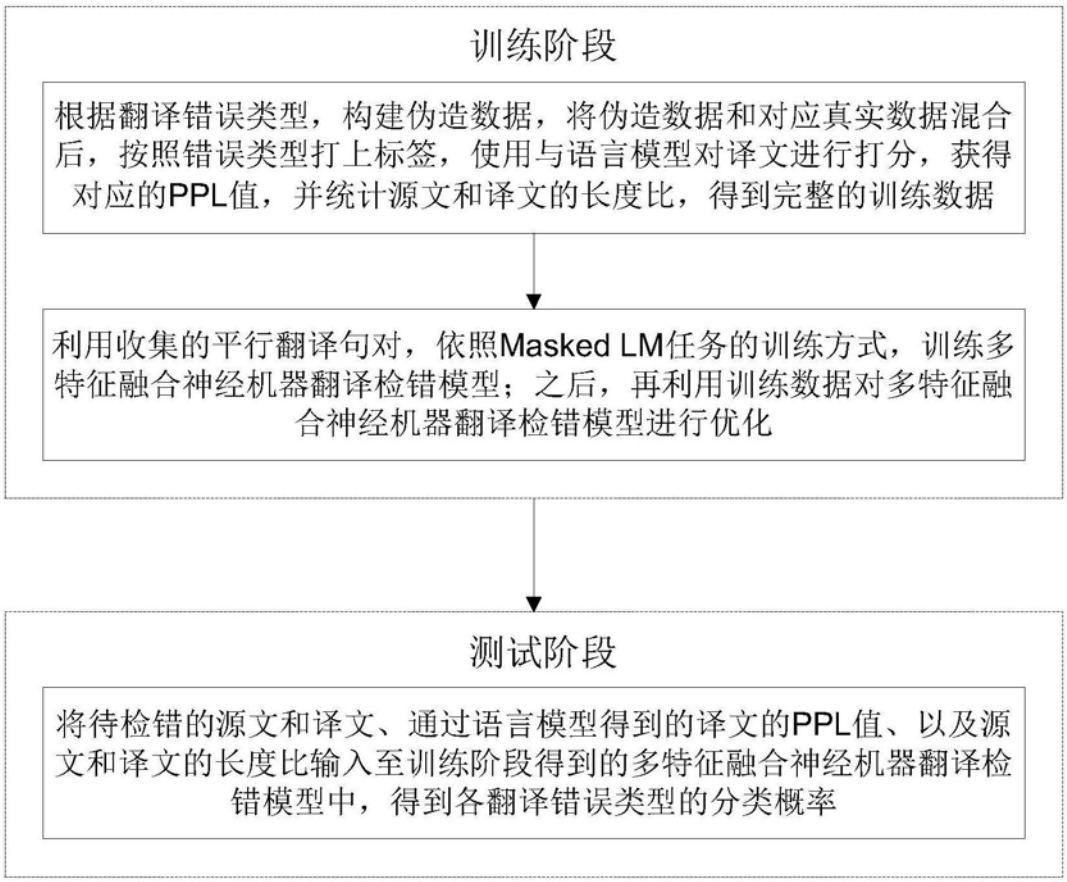

训练阶段
根据翻译错误类型，构建伪造数据，将伪造数据和对应真实数据混合后，按照错误类型打上标签，使用与语言模型对译文进行打分，获得对应的PPL值，并统计源文和译文的长度比，得到完整的训练数据
利用收集的平行翻译句对，依照Masked LM任务的训练方式，训练多特征融合神经机器翻译检错模型；之后，再利用训练数据对多特征融合神经机器翻译检错模型进行优化
测试阶段
将待检错的源文和译文、通过语言模型得到的译文的PPL值、以及源文和译文的长度比输入至训练阶段得到的多特征融合神经机器翻译检错模型中，得到各翻译错误类型的分类概率

图1

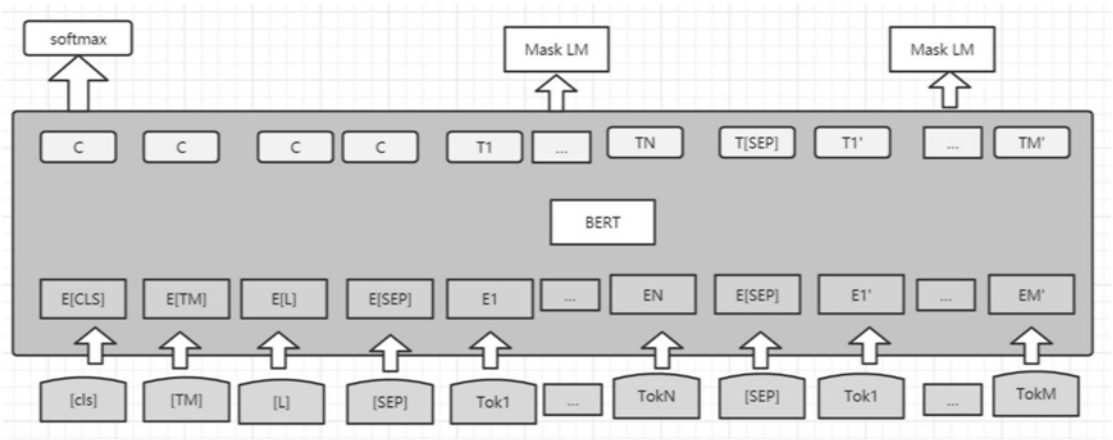

softmax
Mask LM
Mask LM
C
C
C
C
T1
...
TN
T[SEP]
T1'
...
TM'
BERT
E[CLS]
E[TM]
E[L]
E[SEP]
E1
...
EN
E[SEP]
E1'
...
EM'
[cls]
[TM]
[L]
[SEP]
Tok1
...
TokN
[SEP]
Tok1
...
TokM

图2