文章编号：1003-0077（2017）00-0000-00

融入置信度的文本图像翻译研究

伍凌辉[1,2] 马聪[1,2] 韩旭[3,2] 赵阳[4,2] 张亚萍[5,2] 周玉[6,1,2,3]

（1. 中国科学院自动化研究所 模式识别国家重点实验室，北京 100190；
2. 中国科学院大学 人工智能学院，北京 100049；
3. 凡语 AI 研究院 北京中科凡语科技有限公司，北京 100190）

摘要：文本图像翻译旨在将嵌在图像中的源端语言文本翻译成目标语言。文本图像翻译系统通常由相互独立的光学字符识别(Optical Character Recognition, OCR)和机器翻译（Machine Translation, MT）模型级联组成。OCR 模型将文本图像识别成转录文本，MT 模型将转录文本翻译成目标语言。由于 OCR 模型转录文本存在噪声，而 MT 模型对噪声文本表现不佳，文本图像翻译系统性能远不如纯文本机器翻译系统。为缓解噪声文本带来的问题，鲁棒性机器翻译主要采用以下两种方法：1）使用合成噪声文本，以模拟 OCR 转录带来的噪声；2）利用干净文本和噪声文本的对比学习，拉近噪声文本和干净文本的分布。未能考虑以下问题：1）忽视来自 OCR 模型的置信度信息，未能考虑 OCR 和 MT 系统的有效融合；2）仅采用合成噪声，类型单一，无法覆盖实际噪声类型。3）仅采用句子粒度的粗粒度对比损失，忽略细粒度的词的对比信息。为解决这上述问题，该文提出一种融合置信度信息的文本图像翻译方法，充分利用转录文本中每个字符输出的概率分布得到每个词的置信度信息，使级联式文本图像翻译系统中的 OCR 模型和机器翻译模型产生更有效的融合，同时针对 OCR 转录文本的噪声特点，设计一种能提供词粒度的对比信息的监督文本，提升模型性能。实验表明，所提方法在中译英以及英中文本图像翻译任务上相较于传统的管道式模型取得显著的提升。

关键词：置信度；文本图像翻译；鲁棒性神经机器翻译

中图分类号：TP391 **文献标识码：**A

Research of Incorporating Confidence information into Text Image Translation

Wu linghui[1,2] , Ma Cong[1,2] , Han Xu[1,2], Zhao Yang[1,2] ,Zhang Yaping[1,2] ,Zhou Yu[1,2,3]

(1. National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China；2. School of Artificial Intelligence, University of China Academy of Sciences, Beijing 100049, China;3. Fanyu AI Research, Beijing Fanyu Technology Ltd. Beijing 100190, China)

Abstract : Text image machine translation aims to translate the source language embedded in images into the target language. The text image translation system is usually cascaded by optical character recognition (OCR) and machine translation (MT) models. The OCR model recognizes the text image into a transcribed text, and then the MT model translates the transcribed text into the target language. As there are errors in the transcribed text from OCR model and the NMT model is vulnerable to the errors in source transcribed texts from the OCR model, the performance of the text image translation system is far inferior to the text machine translation system. In order to alleviate the problems, the research of robust machine translation mainly adopts the following two methods: 1) using synthetic noise text to simulate the noise caused by OCR transcription; 2) using the contrast learning of clean text and noisy text to narrow the distribution of noisy text and clean text. However, there are two drawbacks in prior work: 1) ignoring confidences information from the OCR model and failing to consider the effective integration of OCR

收稿日期：2022；**定稿日期**：2022
基金项目：国家自然科学基金青年科学基金项目(62106265)

and MT systems; 2) only using synthetic noise, which cannot cover the actual noise type; 3) using the coarse-grained contrastive loss of sentence granularity, ignoring the contrastive information of fine-grained words. To address these issues, we propose a method to incorporate confidence information into text image translation which bridges the gap by reusing the ignored probability distribution of character in OCR to generate confidence for each token. Furthermore, in view of the characteristics of OCR recognition errors, we tailor a supervised text to provide contrastive information with word granularity to improve the systems' performance. Experimental results on Chinese-English and English-Chinese translation tasks demonstrate that our approach achieves significant improvements over the conventional pipeline methods.

Key words: Confidence; Text image translation; Robust NMT

0 引言

文本图像机器翻译旨在将嵌有源语言文本的图像翻译成对应的目标语言。其具有广泛的应用，如翻译扫描或者拍照的文本图像以及烧录在视频中的字幕等等。目前主流方法采用级联管道式[1–3]的方法，如图 1 所示，即先利用 OCR 模型[4–6]从图像中得到转录文本，再利用机器翻译模型[7–10]将转录文本翻译成目标语言。在级联模型中，OCR 模型和机器翻译模型分别独立的进行训练和测试，最后将 OCR 模型和机器翻译模型进行模型级联。然而，独立的级联方式存在误差累积问题，即 OCR 模型识别的噪声文本会导致机器翻译模型性能下降，目前的文本图像机器翻译模型的性能，距离纯文本机器翻译模型仍有很大的距离。

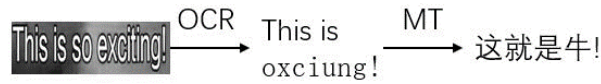


图 1 级联模型示意图

表 1 转录文本噪声示例

文本类型	文本
转录文本	this is so oxciung!
标准文本	this is so exciting!
转录文本(子词)	this is so ox ci ung !
标准文本(子词)	this is so exciting !

表 2 转录文本噪声占比

语言	带噪转录文本	变长转录文本
中文	36.57	30.29
英文	58.99	44.48

鲁棒性神经机器翻译[11–18]致力于缓解噪声文本对机器翻译系统带来的影响，在特定的噪声领域上取得一定的提升，如 ASR 噪声[18]，但将这些方法直接应用到文本图像翻译任务上时存在如下几点不足：

1) 依赖合成文本噪声。现有方法主要采取添加合成文本噪声的方法来提升鲁棒性[15–19]，然而通过 OCR 模型得到的转录文本，其噪声具有难以模拟合成的独特性。如表 1 所示，可以看到，OCR 转录文本中存在的替换、删除、添加等错误，使得输入到机器翻译模型的子词序列长度发生变化。现有的鲁棒性神经机器翻译方法未能考虑这类问题，而这类错误在 OCR 转录文本噪声的中占比非常高，如表 2 所示。

2) 仅使用句子粒度的对比损失。现有的鲁棒性机器翻译方法往往在编码器端使用句子级的粗粒度对比损失[15]，或者利用目标端的翻译损失[15–19]来进行鲁棒性学习。没有在编码器端考虑词粒度的对比信息。

基于上述分析，现有的鲁棒性神经机器翻译方法，未能考虑 OCR 模型和 MT 模型间的有效融合，且存在一定的局限性。本文观察到，利用 OCR 模型输出字符的概率分布得到的置信度信息，可以用于度量识别字符的出错概率。根据本文统计，基于概率计算的置信度其与词错误与否的相关性系数为 0.67，有较强的相关性，同时每个词的置信度越低，其为错词的概率越高。

基于以上观察，本文提出融入置信度信息的文本图像机器翻译方法。通过利用 OCR 输出的概率分布计算每个字符的置信度，进而得到输入到机器翻译模型中每个词的置信度。具体地，本文提出一种融入置信度的注意力模块，并将其应用在编码器中。为使得 OCR 模型和机器翻译模型能更有效的融合，本文利用 OCR 转录文本来训练含

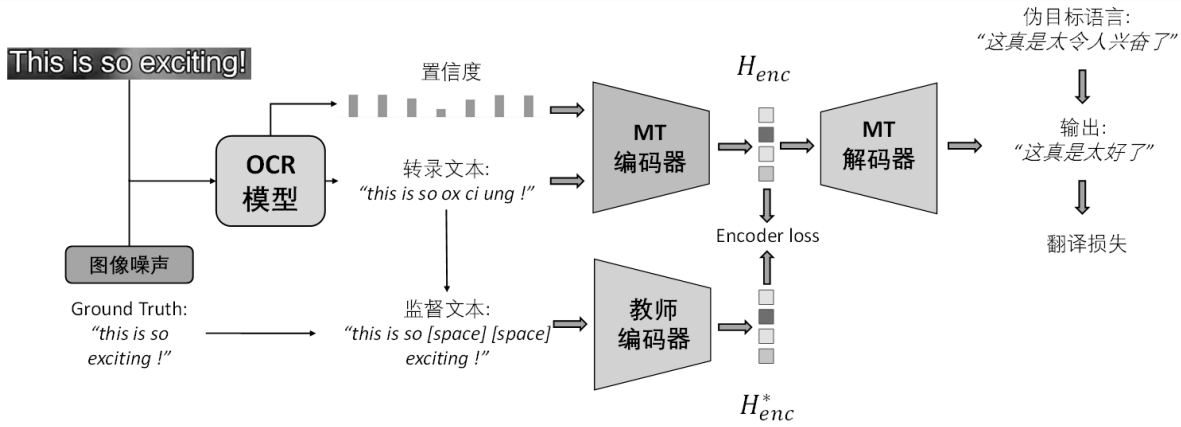


图 2 模型结构示意图

有融入置信度的注意力模块的 MT 编码器模型，并针对 OCR 的噪声特点，设计对应的监督文本，该监督文本使得进行词粒度的监督学习成为可能，实现在不增加额外数据的情形下，仅利用原有的 OCR 和机器翻译数据集来提升级联管道式系统的翻译性能。

综上所述，本文的贡献主要分为以下三点:

1) 提出一种新颖的融入置信度信息的文本图像翻译方法，实现级联式文本图像翻译系统中的 OCR 和 MT 模型的有效融合。

2) 针对 OCR 噪声特点，设计一种能提供词级别对比信息的监督文本，细粒度指导鲁棒性机器翻译模型编码器的更新。

3) 实验表明，在不增加额外数据的情形下，仅利用原有的 OCR 和机器翻译数据集，融入置信度信息的文本图像机器翻译方法能有效提升级联管道式模型的翻译性能。

1 融入置信度的文本图像翻译方法

文本图像翻译旨在将嵌入在文本图像中的源端语言文本翻译成目标语言。当前的方法采用管道式的方法，即利用 OCR 训练数据集和机器翻译训练数据集各自训练一个 OCR 模型和机器翻译模型。在生成阶段，OCR 模型将图像识别成转录文本，机器翻译模型再将转录文本翻译成目标语言。本文旨在不增加额外数据的情形下，即仅利用 OCR 训练数据集和机器翻译训练数据集来提升级联管道式模型的整体翻译效果。

1.1 模型架构

在本文所提出的方法中，其模型整体结构如图 2 所示，模型主要由四个模块组成：1）OCR 模型；2）机器翻译模型，其中包括教师编码器和 MT 解码器；3）MT 编码器，其含有融入置信度的注意力模块。

训练分两个阶段，在第一阶段，通过利用 OCR 训练集训练 OCR 模型，利用机器翻译训练集训练机器翻译模型，即教师编码器和 MT 解码器，使得教师编码器能够在下一阶段的训练中指导 MT 编码器的参数更新。

第二阶段，训练过程如图 2 所示，通过利用 OCR 转录文本和对应的置信度训练 MT 编码器，以实现 OCR 模型和机器翻译模型的有效融合，在此训练过程中其余模块的参数均被固定。

在第二阶段的训练过程中，主要分为以下四个步骤：

1) 转录文本的生成：首先通过 OCR 模型得到转录文本以及转录文本中每个字符的概率分布。

2) 置信度计算：得到每个字符的概率分布后，通过计算得到每个字符的置信度，进而得到每个词的置信度。

3) 融入置信度的注意力模块：在编码器的编码过程中，为减少噪声信息的影响，提出一种融入置信度的注意力机制。

4) 监督文本的生成：由于第二训练阶段使用的数据是通过 OCR 模型生成转录文本，其缺乏真实的目标语言，为了让编码器能有效地训练，本文通过利用转录文本对应的标准文本生成监督文本用来提供词级别对比信息。

接下来，将详细介绍第二训练阶段的主要步骤以及模型的推理阶段。

1.1.1 转录文本生成

本文利用 OCR 训练数据集和训练好的 OCR 模型生成转录文本，为提升转录文本的错误率，通过在将 OCR 训练数据输入 OCR 模型前在图像

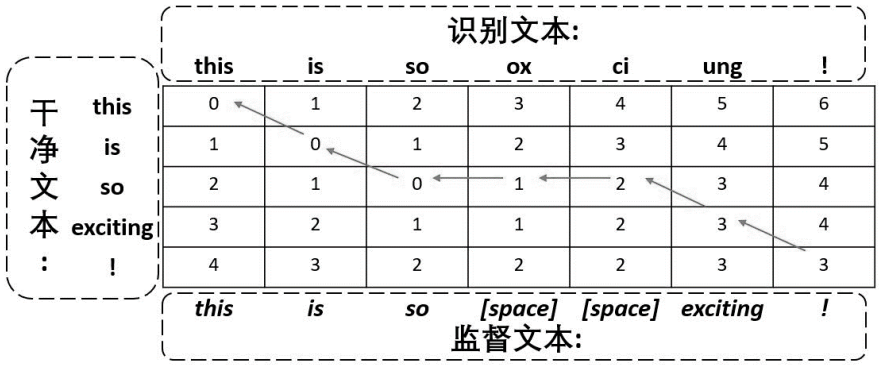


图 3 监督文本生成示意图

上分别添加椒盐噪声和高斯噪声以增加转录文本中的错误率。

1.1.2 置信度计算

对于计算 OCR 模型对所识别字符的置信度，遵循以下的原则：若 OCR 模型给出的这个字符的概率分布就越集中，那么它对这个字符的置信度越高。因此，给定一个字符的输出概率分布$P(y|x_{img},\theta_{ocr})$，本文采取了三种计算置信度的方法：

1） 基于概率的字符置信度：

$$C_p = \max P\left(y|x_{img},\theta_{ocr}\right) \tag{1}$$

2） 基于方差的字符置信度：

$$C_v = \sum_{i}^{N}(P\left(y_i|x_{img},\theta_{ocr}\right) - \frac{1}{N})^2 \tag{2}$$

3） 基于熵的字符置信度：

$$C_e = E_{MAX} + \sum_{i}^{N} P\left(y_i|x_{img},\theta_{ocr}\right)\log \mathrm{P}\left(y_i|x_{img},\theta_{ocr}\right) \tag{3}$$

其中，E_{MAX}表示熵的最大值。为将置信度设置在 0-1 之间，对每一种计算方法得到的置信度，都进行归一化。

由于输入到机器翻译模型中的文本是以词为单位的序列，为得到每一个词的置信度，假设所有字符概率分布条件独立，因此，可以通过如下公式得到词的置信度。

1) 基于概率的词置信度：

$$C_p(token) = \prod_{\hat{y}\in token} C_p(\hat{y}) \tag{4}$$

2) 基于方差的词置信度：

$$C_v(token) = \prod_{\hat{y}\in token} C_v(\hat{y}) \tag{5}$$

3) 基于熵的词置信度：

$$C_e(token) = \frac{1}{n}\sum_{\hat{y}\in token} C_e(\hat{y}) \tag{6}$$

其中n表示 token 的字符数。

1.1.3 融入置信度的注意力模块

为能够让 MT 编码器利用置信度信息更好的编码转录文本，本文设计一个融入置信度的注意力模块。在该注意力模块下，模型编码遵循如下原则：若一个词的置信度较低，那么就更少的参考这个词的信息，若这个词的置信度较高，那么就更多的参考这个词的信息。形式地来说，对于每一句转录文本，有其对应的置信度向量$C \in R^{1\times L}$，那么本文将原有的注意力公式修改为如下形式：

$$Attention(Q,K,V,C) = softmax\left(\frac{QK^T}{\sqrt{D}}\right)(C\odot V) \tag{7}$$

从公式中可以看到，对于置信度较低的词，其表示会更少的参与到下一层的表示的生成，而置信度较高的词会更多的参与到下一层表示的生成。

1.1.4 监督文本的生成

本文训练的目标是希望 MT 编码器能在利用置信度信息的基础上将转录文本编码成更好的隐层状态表示以供 MT 解码器解码。

但正如引言中所述，OCR 模型识别是以字符为单位的识别，而机器翻译模型的输入是以 BPE[20]分词后的子词为单位的序列。因此单个字符的错误会引起分词出错，进而改变序列的长度，使得难以定位每个词的标准答案。为能够提供词粒度的监督信号，本文提出通过利用标准文本和转录文本的编辑距离矩阵来生成相应的监督文本的方法。如图 3 所示，首先通过计算得到编辑距离矩阵，之后在编辑距离矩阵上回溯得到转录文本相应的监督文本。该监督文本如下两点好处：

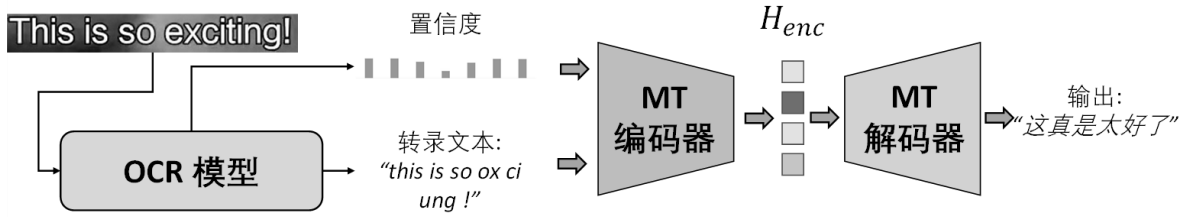


图 4 模型推理结构示意图

1) 它和转录文本有着相同的文本序列长度。
2) 在文本序列长度相同的情况下，它与标准文本间的编辑距离最小。

同时，在中译英和英译中两个翻译任务上测试监督文本的翻译效果，如表 3 所示：

表 3 监督文本翻译效果比对

	中译英	英译中
转录文本	16.59	22.59
标准文本	19.14	26.28
监督文本	19.03	25.59

因此，本文认为该监督文本是一种好的监督信号。

1.2 模型的训练与推理

模型采用反向传播算法进行训练，模型的损失函数主要来源于两个部分。一是来源于 MT 编码器编码转录文本得到的隐层状态表示H_{enc}和教师编码器编码监督文本得到的隐层状态表示H^*_{enc}距离损失L_{dis}；二是将H_{enc}输入到 MT 解码器中的翻译损失L_{trans}，(由于没有标准的目标语言，因此本文预先通过预训练的机器翻译模型，即教师编码器和 MT 解码器，得到标准文本对应的翻译，从而构建的伪平行语料)。

为度量隐状态表示的距离损失，本文采取 Huber loss 函数度量距离损失L_{dis}，Huber loss 函数如下：

$$L_\delta = \begin{cases} \frac{1}{2}(y-\hat{y})^2 & if\ |y-\hat{y}| < \delta \\ \delta\left(|y-\hat{y}| - \frac{1}{2}\delta\right) & otherwise \end{cases} \quad (8)$$

而翻译损失则由交叉熵来度量。最终损失为它们的加和形式。具体公式如下：

$$L = L_{dis} + L_{trans}$$

在推理阶段，模型结构如图 4 所示，模型仅保留 OCR 模型、由训练第二阶段得到的 MT 编码器模型以及 MT 解码器模型。在推理阶段中，图片输入到 OCR 模型中得到转录文本和对应的置信度，之后将置信度和转录文本输入到 MT 编码器中编码得到隐层状态表示H_{enc}，最后将H_{enc}输入到 MT 解码器中得到最终的翻译。

2 实验

2.1 数据构建

鉴于目前 OCR 领域中的开源的数据集如 MJSynth[21]和 SynthText[22]等都是以词为单位的，不适用于翻译任务。因此，本文为中文文本识别任务构建六百万的训练数据，为英文文本识别任务构建一百万的训练数据。不同于之前的数据集，该数据中每一张图片含有一个句子的文本而不是一个单词。该数据是通过开源的文本识别数据生成工具[1]生成的。其文本数据来源于 WMT18[2]中英翻译任务，背景图片来源于 Youtube[3]。由于文本图像中的文本往往较短，因此本文用句子字符长度对文本数据进行过滤。其中，英文的过滤长度为 80，中文为 40。需要指出的是，本文在后续的实验中并未使用平行语料，而是仅仅使用其源端语言来构建 OCR 训练数据集,这是由于在实践中 OCR 训练集和 MT 训练集往往没有关联性。

为评估本文所提出的方法，本文分别构造合成测试集和字幕测试集。合成测试集的文本数据来源于 WMT 和 IWSLT 的测试数据集，其过滤方法与训练集相同。字幕测试集则是通过从视频中收集双语字幕构建的。合成测试集的大小是 2502 句，字幕测试集的大小为 1040 句。

[1] https://github.com/Belval/TextRecognitionDataGenerator
[2] http://www.statmt.org/wmt18/
[3] https://www.youtube.com/

2.2 实验设置

为验证所介绍方法的有效性，本文在中译英和英译中两个翻译任务进行了实验，并采取 4-gram BLEU[23]来评估 MT 翻译质量。采用 CER[24]（字符错误率，Character Error Rate），评估 OCR 转录质量。

本文采用扩充的 NIST 数据集和合成的 OCR 训练集作为 OCR 模型和机器翻译模型的训练集。对于机器翻译的训练集，本文通过 sentencepiece[4]来进行字节对编码（Byte-pair encoding, BPE）[20]，中英的词表大小均设置为 32768。

在训练 MT 编码器时，本文仅从 OCR 训练集中取前一百万的数据作为训练集。本文从合成测试集中选取前 1k 句作为验证集。

模型方面，本文的基线模型有两个模型组成：

1) OCR 模型：本文采用 OCR 模型是基于 Baek[25]等人所开源的工作[5]。本文所实现的 OCR 模型设置与 Baek[25]论文的最优模型一致，但由于图片所含文本长度不同，因此将处理后的图片高度设置为 32，宽度则与识别语种相关，中文图片宽度设置为 160，英文则设置为 320。

2) 机器翻译模型：本文采用 Transformer 作为机器翻译模型。参数设置与 Vaswani 等人[10]基本一致，仅将$p_{dropout}$设置为 0.3。

在第二训练阶段中的 MT 编码器模型与基线模型的编码器模型结构相同。在训练前，本文用教师编码器参数来初始化其参数。

2.3 对比实验设置

为验证本文提出的方法，除基础的基线模型以外，本文对比两种鲁棒性学习的方法和一种 OCR 后校正方法。为公平比较，所有方法的模型结构都与基线模型相同。方法如下：

1) 词替换：Wang 等人[26]提出通过用混淆词表中的其他词随机替换源语言句子和目标语言句子中的词来增强训练数据，提升模型的鲁棒性。在本文的实现中，由于构建混淆词表耗时巨大，不同的 OCR 模型，混淆词表不同，因此采用全词表作为每一个词的混淆词表，同时仅替换源语言句子中的词，替换概率设置为 0.1。

2) 对抗学习：Miyato 等人[27]提出利用对抗性学习来提升文本分类的鲁棒性。其实现方法是在词向量上添加扰动来构建对抗样本，从而实现对抗学习。本文将其应用到机器翻译中来作为对比。

3) OCR 后校正：OCR 后校正即在 OCR 识别后通过一个校正模块对转录文本进行纠错。本文采用 Pycorrector[6] 来进行纠错，再进行翻译。

2.4 实验结果

表 4 合成测试集实验结果

方法	中译英		英译中	
	Valid	Test	Valid	Test
标准文本翻译结果				
基线模型[10]	17.98	19.14	22.87	26.28
管道模型翻译结果				
基线模型[10]	15.59	16.59	20.27	22.59
词替换[26]	14.79	16.46	20.40	22.55
对抗学习[27]	14.80	16.48	20.08	22.65
OCR 后校正 [6]	15.52	16.59	20.11	21.89
融入置信度的翻译结果				
$+C_p$	16.79	17.62	**21.33**	23.34
$+C_d$	16.74	17.70	21.26	**23.36**
$+C_e$	**16.89**	**17.76**	21.02	23.10

表 5 字幕测试集实验结果

方法	中译英	英译中
标准文本翻译结果		
基线模型[10]	14.52	20.30
管道模型翻译结果		
基线模型[10]	13.71	19.29
词替换[26]	13.16	19.69
对抗学习[27]	13.48	19.52
OCR 后校正[6]	13.68	19.62
融入置信度的翻译结果		
$+C_p$	14.49	20.53
$+C_v$	14.54	20.50
$+C_e$	**14.60**	**20.54**

表 6 转录文本和校正文本 CER

	中文合成测试集	中文字幕测试集	英文合成测试集	英文字幕测试集
转录文本	12.42	5.00	5.44	7.34
校正文本	12.67	5.02	6.11	3.59

表 4 和表 5 分别是在合成测试集和字幕测试

[4] https://github.com/google/sentencepiece
[5] https://github.com/clovaai/deep-text-recognition-benchmark
[6] https://github.com/shibing624/pycorrector

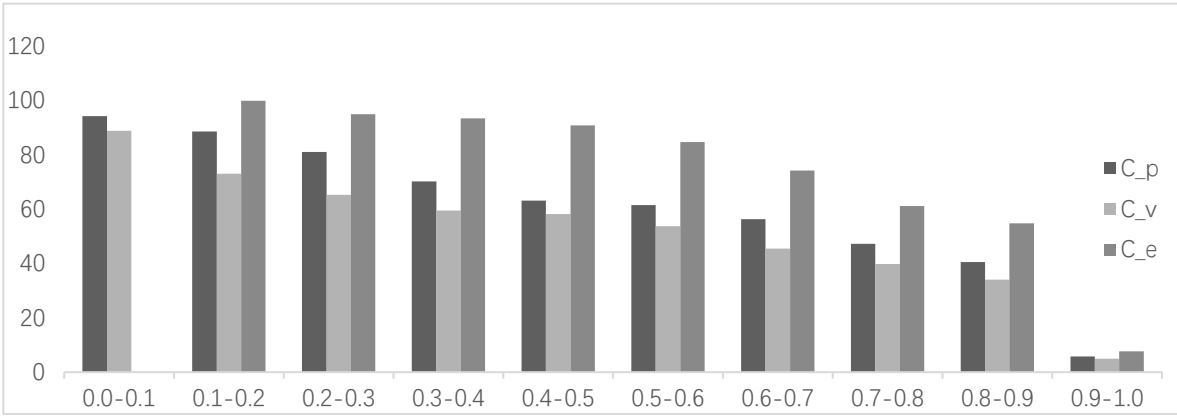


图 5 不同置信度下的词错误率

集上的实验结果。在合成测试集上，融入置信度信息后，本文在英译中和中译英任务上分别取得 0.77BLEU 和 1.17BLEU 的提升。在字幕测试集上，可以看到其表现甚至好于基线模型在标准文本的翻译结果。这是由于在 MT 编码器时所用的转录文本长度分布与字幕数据更一致，因此模型能更好编码从字幕中得到的转录文本。

同时，可以看到，仅进行词替换或者在词向量上引入扰动，难以提升翻译模型在 OCR 转录文本上的性能。其原因如表 1 中所示的，转录文本中的噪声无法被的词替换型的错误可以涵盖的。

针对使用 OCR 后校正后，翻译性能反而下降的问题，本文统计 OCR 转录文本及其校正文本的 CER，如表 6 所示，发现采用校正模型后，在三个数据集，反而导致了 CER 的增高，仅在英文字幕测试集上出现了 CER 的下降。从翻译结果来看，应用校正模型也只提升在英译中的字幕测试集上的性能，在其他测试集上性能均有下降。说明引入普通校正模型容易引入校正错误，这点与 Amrhein 等人[28]在文章中指出的相一致。同时也说明通用型的校正模型难以直接应用到文本图像翻译任务中，因此在实际应用中需针对 OCR 模型单独训练校正模型以对转录文本进行校正提高校正准确率。而我们的方法并不需要校正模型就能提升总体的翻译性能。

2.5 实验分析

2.5.1 消融性实验

为进一步探究不同的损失函数和置信度的影响，本文在中译英任务上进行消融性实验。实验结果如表 7 所示。C表示使用置信度，在消融性实验中选取的是基于方差的置信度C_v，表格中的 1 和 0 表示是否在实验中使用置信度或损失函数，1 表示使用，0 表示未使用。

表 7 消融性实验结果

C	L_{trans}	L_{dis}	合成测试集		字幕测试集
			Valid	Test	Test
0	0	1	15.46	16.60	14.40
0	1	0	16.18	17.15	13.31
1	0	1	16.30	17.36	14.51
1	1	0	16.05	16.81	13.69
1	1	1	**16.74**	**17.70**	**14.54**

从表中，可以看出仅使用L_{trans}或L_{dis}无法在两个测试集上取得一致的提升。如仅加入L_{trans}，虽然在合成测试集取得较高的提升，但在字幕测试集上有较大的下滑，而引入置信度信息后，虽然相较仅加入L_{trans}在合成测试集上翻译效果有一定的下滑，但在字幕测试集上相较基线模型没有出现明显下滑，可见引入置信度信息后明显提升模型的泛化性能。而加入置信度和L_{dis}比加入置信度和L_{trans}翻译效果更好的原因是前者不会损害模型在没有噪声的转录文本上的性能。这一点在文本图像翻译任务中非常重要，因为实践有很多 OCR 模型识别得到的转录文本是没有噪声的。最后，可以看到通过加入置信度和两种损失函数，模型在两个测试集上都取得最好的翻译结果，证明了所提出方法的有效性。

2.5.2 置信度分析

为探究三种置信度信息的异同，统计中文中三种置信度在不同置信度区间下的词错误率，如图 5 所示。可以看到随着置信度的不断上升，在三种置信度上，词错误率都随之上升。因此三种置信度都能在一定程度上判断词错误的概率，而基于熵的置信度几乎在所有区间的词错误率都高于其他置信度。这是很可能是因为基于熵的置信度较少的给予转录正确词较低的置信度，即置信度误判较少。同时这也说明，引入置信度信息虽然能屏蔽错误文本的信息，但一旦置信度信息误判较多，也会更多的屏蔽正确文本信息，也会对

翻译性能造成影响。因此探索更加精准的置信度度量方法有着非常深远的意义。

2.5.3 翻译结果分析

如本文在引言中所说,OCR 转录文本中噪声会引起输入机器翻译模型的文本序列长度发生变化。而以往的方法难以在此类型的噪声上奏效。因此，本文在中译英的合成测试集上统计不同方法在变长文本上的翻译结果，如表 8 所示，从表格中可以看到以往的方法在应对变长文本上均难以奏效，而本文提出的方法在变长文本上均取得超过 1 个 BLEU 的显著提升。

表 8 变长文本上的翻译结果

方法	变长文本	$\triangle BLEU$
管道模型翻译结果		
基线模型[10]	15.05	0
词替换[26]	14.93	-0.12
对抗学习[27]	14.59	-0.46
OCR 后校正 [6]	15.07	0.02
融入置信度的翻译结果		
$+C_p$	16.18	+1.13
$+C_v$	16.31	+1.26
$+C_e$	16.29	+1.24

3 相关工作

本文的工作主要与四个方面的研究密切相关: 1）文本识别；2）鲁棒性神经机器翻译；3）置信度度量；4）OCR 后校正。

1) **文本识别**。随着神经网络的发展，He 等人[4]、Shi 等人[5]和 Su 等人[29]提出利用卷积神经网络和循环神经网络来编码图片信息，并使用连接时序分类[30]（Connectionist Temporal Classification, CTC）进行解码。Shi 等人[31]和 Cheng 等人[32]提出基于注意力机制的解码方法。Baek 等人[25]总结前人的工作，提出一个统一的文本识别框架。Qian 等人[33]则提出利用语义信息帮助模型进行文本识别。Fang 等人[6]则提出引入语言模型来提升识别效果。本文的工作选取 Baek 等人[25]中的最优模型作为本文的 OCR 模型，但本文所提出的方法也可以应用到其他的 OCR 模型中。

2) **鲁棒性神经机器翻译**。神经机器翻译模型在带噪文本上的翻译性能日益受到关注。Belinkov 等人[13]和 Karpukhin 等人[14]提出利用合成噪声和自然噪声提升模型的鲁棒性。Cheng 等人[15]、Ebrahimi 等人[16]、Cheng 等人[17]提出在词向量中添加扰动来自动地构建对抗样本以提升模型的鲁棒性。Xue 等人[18]提出利用拼音信息以提升模型在带有语音识别噪声文本上的鲁棒性。Michel 等人[34]提出一套用于评估机器翻译鲁棒性的数据集，但此数据集中的噪声文本来源于互联网，而非 OCR 模型。与前人的工作不同的是，本文不是简单的通过合成文本噪声来构建对抗样本。针对文本图像翻译任务的特点，本文提出利用置信度的信息来减少噪声的影响，同时针对 OCR 转录文本噪声的特点，设计细粒度的损失函数，进而提升模型在 OCR 转录文本上的鲁棒性。

3) **置信度度量**。在置信度的度量有很多的工作。Rikters 等人[35]提出利用注意力分布作为度量置信度的依据。Wang 等人[36]使用模型的不确定性来作为度量置信度的依据。Kim 等人[37]提出一种两阶段的神经网络估计模型来估计模型的输出质量。与前人的工作不同，本文仅仅只是利用 OCR 模型输出的概率分布来度量置信度，未来将探索更多的置信度度量方法。

4) **OCR 后校正**。由于 OCR 中的噪声对下游任务的性能有很大的影响，OCR 后校正也越来越受到重视。Tong 等人[38]、Kolak 等人[39]利用统计语言模型来进行校正。近期也有不少利用神经网络的工作，Nastase 等人[40]提出利用序列到序列的模型来进行纠错。Neudecker 等人[41]提出两阶段的方法来自动进行 OCR 后校正。在本文的实验中，仅和 Pycorrect[6] 进行对比。这是考虑到其实现简单，同时在现实中被广泛利用。

4 结论

本文针对文本图像翻译任务，实现级联式文本图像翻译系统中的 OCR 和 MT 模型的有效融合。同时针对 OCR 转录文本的噪声特点，设计能提供词粒度的对比信息的监督文本，提升模型性能。实验表明，所提方法在中英以及英中文本图像翻译任务上相较于传统的管道式模型，取得显著的提升。

参考文献

[1] Watanabe Y, Okada Y, Kim Y B, et al. Translation camera[C]//Proceedings. Fourteenth International Conference on Pattern Recognition (Cat. No. 98EX170). IEEE, 1998, 1: 613-617.

[2] Yang J, Chen X, Zhang J, et al. Automatic detection and translation of text from natural scenes[C]//2002 IEEE International conference on acoustics, speech, and signal processing. IEEE, 2002, 2: II-2101-II-2104.

[3] Hinami R, Ishiwatari S, Yasuda K, et al. Towards fully automated manga translation[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2021, 35(14): 12998-13008.

[4] He P, Huang W, Qiao Y, et al. Reading scene text in deep convolutional sequences[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2016: 3501-3508

[5] Shi B, Bai X, Yao C. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition[J]. IEEE transactions on pattern analysis and machine intelligence, 2016, 39(11): 2298-2304.

[6] Fang S, Xie H, Wang Y, et al. Read like humans: autonomous, bidirectional and iterative language modeling for scene text recognition[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 7098-7107.

[7] Cho K, Van Merriënboer B, Gulcehre C, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation[J]. arXiv preprint arXiv:1406.1078, 2014.

[8] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate[J]. arXiv preprint arXiv:1409.0473, 2014.

[9] Gehring J, Auli M, Grangier D, et al. Convolutional sequence to sequence learning[C]//International Conference on Machine Learning. PMLR, 2017: 1243-1252.

[10] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//Advances in neural information processing systems. 2017: 5998-6008.

[11] Szegedy C, Zaremba W, Sutskever I, et al. Intriguing properties of neural networks[C]//2nd International Conference on Learning Representations, ICLR 2014. 2014.

[12] Goodfellow I J, Shlens J, Szegedy C. Explaining and harnessing adversarial examples [J]. stat, 2015, 1050: 20.

[13] Belinkov Y, Bisk Y. Synthetic and natural noise both break neural machine translation[J]. arXiv preprint arXiv:1711.02173, 2017.

[14] Karpukhin V, Levy O, Eisenstein J, et al. Training on synthetic noise improves robustness to natural noise in machine translation[J]. W-NUT 2019, 2019: 42.

[15] Cheng Y, Tu Z, Meng F, et al. Towards robust neural machine translation[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2018: 1756-1766.

[16] Ebrahimi J, Lowd D, Dou D. On adversarial examples for character-level neural machine translation[C]//Proceedings of the 27th International Conference on Computational Linguistics. 2018: 653-663.

[17] Cheng Y, Jiang L, Macherey W. Robust neural machine translation with doubly adversarial inputs[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019: 4324-4333.

[18] Xue H, Feng Y, Gu S, et al. Robust neural machine translation with ASR errors[C]//Proceedings of the First Workshop on Automatic Simultaneous Translation. 2020: 15-23.

[19] Liu H, Ma M, Huang L, et al. Robust neural machine translation with joint textual and phonetic embedding[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019: 3044-3049.

[20] Sennrich R, Haddow B, Birch A. Neural machine translation of rare words with subword units[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2016: 1715-1725.

[21] Jaderberg M, Simonyan K, Vedaldi A, et al. Synthetic data and artificial neural networks for natural scene text recognition[J]. arXiv preprint arXiv:1406.2227, 2014.

[22] Gupta A, Vedaldi A, Zisserman A. Synthetic data for text localisation in natural images[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 2315-2324.

[23] Papineni K, Roukos S, Ward T, et al. Bleu: a method for automatic evaluation of machine translation[C]//Proceedings of the 40th annual meeting of the Association for Computational Linguistics. 2002: 311-318.

[24] Sueiras J, Ruiz V, Sanchez A, et al. Offline continuous handwriting recognition using sequence to sequence neural networks[J]. Neurocomputing, 2018, 289: 119-128.

[25] Baek J, Kim G, Lee J, et al. What is wrong with scene text recognition model comparisons? dataset and model analysis[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 4715-4723.

[26] Wang X, Pham H, Dai Z, et al. SwitchOut: an efficient data augmentation algorithm for neural machine translation[C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. 2018: 856-861.

[27] Miyato T, Dai A M, Goodfellow I et al. Adversarial training methods for semi-supervised text classification[J]. 2017(2015): 1–10.

[28] Amrhein C, Clematide S. Supervised OCR error detection and correction using statistical and neural machine translation methods[J]. Journal for Language Technology and Computational Linguistics (JLCL), 2018, 33(1): 49-76.

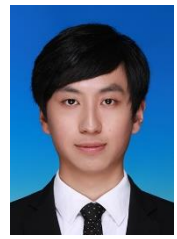
[29] Su B, Lu S. Accurate recognition of words in scenes without character segmentation using recurrent neural network[J]. Pattern Recognition, 2017, 63: 397-405.

[30] Graves A, Fernández S, Gomez F, et al. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks[C]//Proceedings of the 23rd international conference on Machine learning. 2006: 369-376.

[31] Shi B, Wang X, Lyu P, et al. Robust scene text

recognition with automatic rectification[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 4168-4176.
[32] Cheng Z, Bai F, Xu Y, et al. Focusing attention: towards accurate text recognition in natural images[C]//Proceedings of the IEEE international conference on computer vision. 2017: 5076-5084.
[33] Qiao Z, Zhou Y, Yang D, et al. Seed: Semantics enhanced encoder-decoder framework for scene text recognition[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 13528-13537.
[34] Michel P, Neubig G. MTNT: A testbed for machine translation of noisy text[C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. 2018: 543-553.
[35] Rikters M, Fishel M. Confidence through attention[J]. arXiv preprint arXiv:1710.03743, 2017.
[36] Wang S, Liu Y, Wang C, et al. Improving back-translation with uncertainty-based confidence estimation[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). 2019: 791-802.
[37] Kim H, Lee J H, Na S H. Predictor-estimator using multilevel task learning with stack propagation for neural quality estimation[C]//Proceedings of the Second Conference on Machine Translation. 2017: 562-568.
[38] TONG X. A statistical approach to automatic OCR error correction in context[C]//Proceedings of the Fourth Workshop on Very Large Corpora, 1996, Copenhagen, Denmark. 1996: 88-100.
[39] Park Y A, Levy R. Automated whole sentence grammar correction using a noisy channel model[C]//Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. 2011: 934-944.
[40] Nastase V, Hitschler J. Correction of OCR word segmentation errors in articles from the ACL collection through neural machine translation methods[C]//Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). 2018: 706-711
[41] Schaefer R, Neudecker C. A two-step approach for automatic OCR post-correction[C]//Proceedings of the The 4th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature. 2020: 52-57.

伍凌辉（1996—），硕士研究生，学生，硕士研究生，主要研究领域为自然语言处理和机器翻译
E-mail：linghui.wu@nlpr.ia.ac.cn



马聪（出生年—），博士研究生，学生，主要研究领域为自然语言处理和机器翻译。
E-mail：cong.ma@nlpr.ia.ac.cn

周玉（出生年—），通信作者，博士，研究员， 主要研究领域为自然语言处理和机器翻译。
E-mail：yzhou@nlpr.ia.ac.