基于伪数据的机器翻译质量估计模型的训练

吴焕钦[1,2] 张红阳[1] 李静梅[2] 朱俊国[1] 杨沐昀[1,†] 李生[1]

1. 哈尔滨工业大学计算机科学与技术学院，哈尔滨 150001；2. 哈尔滨工程大学计算机科学与技术学院，哈尔滨 150001；

† 通信作者，E-mail: yangmuyun@hit.edu.cn

摘要 为向基于深度学习的机器翻译质量估计模型提供高效的训练数据，提出面向目标数据集的伪数据构造方法，采用基于伪数据预训练与模型精调相结合的两阶段模型训练方法对模型进行训练；并针对不同伪数据规模进行了实验设计。实验结果表明，在构造得到的伪数据下利用两阶段训练方法训练得到的机器翻译质量估计模型，其给出的得分与人工评分的相关性有了显著的提升。

关键词 机器翻译质量估计；深度学习；伪数据

中图分类号

Training Machine Translation Quality Estimation Model Based on Pseudo Data

WU Huanqin[1,2], ZHANG Hongyang[1], LI Jingmei[2], ZHU Junguo[1], YANG Muyun[1] LI Sheng[1]

1. Computer Science and Technology, Harbin Institute of Technology, Harbin, 150001; 2. Computer Science and Technology, Harbin Engineering University, Harbin, 150001

Abstract Aimed at providing efficient training data for neural translation quality estimation model, a pseudo data construction method for target dataset is proposed; the model is trained by two stage model training method: pre training based on pseudo data and fine tuning ; The experimental design of different pseudo data scale is carried out. The experiment results show that the Machine Translation quality estimation model trained by the pseudo data has significantly improved in the correlation between the scores given by human and the artificial scores.

Key words machine translation quality estimation ; deep learning ; pseudo data

在机器翻译研究中，质量估计（Quality Estimation, 简称 QE）是指对机器翻译系统的输出进行质量预测，其结果可以快速的判断出机器翻译质量的好坏，对机器翻译性能的提升起着指导作用[2]。与机器翻译评价不同的是，QE 系统的目标是在不依赖于参考译文的情况下对机器翻译系统的输出进行质量预测。在 QE 任务中，句子级的 QE 是最为流行的研究内容，其任务是为机器译文的每个句子估计一个得分。早期的质量估计研究，主要是利用机器学习的方法，将其看作是一个回归或分类问题，实现对机器翻译译文质量的估计，其主要的研究内容在于特征的抽取及选择。

随着深度学习的引入，通过深度神经网络强大的特征学习能力，解决了以往需要人工设计特征的重要难题，例如，在机器翻译研究中，基于双向长短时记忆神经网络（Long Short Term Memory，简称 LSTM）的序列到序列（Sequence to Sequence）的建模方法取得了显著的效果[1]。近些年来，在机器翻译质量估计研究中，也出现了许多基于深度学习的机器翻译质量估计方法。Zhu J 等[2]提出通过学习双语句子的特征表示来建立机器翻译质量估计模型，Kim H 等[3]提出通过循环神经网络对机器翻译质量估计进行建模，Kreutzer J[4]等提出使用神经网络结构实现对词语级的机器翻译质量估计。实验结果表明，这些基于深度学习的质量估计模型在不使用大量复杂的语言学特征的情况下，仍能取得可观的效果。对于基于深度学习的翻译质量估计模型的训练而言，需要大量的具有不同得分的机器译文对模型进行训练。在目前已有的训练方法中，Zhu 等使用了数据规模较大的双语平行语料让模型对翻译过程中的正例（即翻译完全正确）与反

资助 国家高技术研究发展计划（863 计划） NO.2015AA015405；国家自然科学基金 NO.61370170；NO.61402134

收稿日期:；修回日期: ；网络出版时间：

网络出版地址：

例（即翻译完全错误）的情况进行学习，在一定程度上缓解了训练语料不足的问题，但该方法只是将机器译文简单的分为翻译正确和错误两种情况，并没有对翻译过程中错误的具体信息进行刻画与学习。Kim H等[3]通过大规模的双语平行语料训练得到神经机器翻译系统，在神经机器翻译系统上抽取权重信息，在此基础上，利用质量估计评测任务提供的 11271 句带机器译文得分的训练数据继续训练神经网络，该方法仅使用了少量的具有不同得分的机器译文对模型进行训练，难以学习得到性能较好的神经网络。

鉴于上述方法存在的缺陷，本文提出面向目标数据集的伪数据构造方法，即针对机器翻译过程中可能存在翻译错误的情况构造机器译文，以解决现有训练方法中缺乏具有不同得分的训练数据的问题。，此外，采用基于伪数据的预训练与模型精调的两阶段训练方法对机器翻译质量估计模型进行训练。实验结果表明，本文提出的方法具有较好的性能。

1 基于双向 LSTM 的机器翻译质量估计模型

双向 LSTM 以其强大的序列到序列的建模能力，在机器翻译研究领域取得了巨大的成功，目前基于双向 LSTM 的神经机器翻译系统的性能已取得比传统统计机器翻译系统更好的性能。此外，在已有的基于深度学习的机器翻译质量估计的方法中，基于双向 LSTM 的建模方法也被广泛的应用，例如，Zhu、Kim 等都采用了该方法对机器翻译质量估计进行建模。

因此，本文选取了由 Zhu 等提出的具有一般性的基于双向 LSTM 的机器翻译质量估计模型，通过对该模型的训练以验证伪数据的性能。本文所选取的机器翻译质量估计模型的结构图如图 1 所示。该模型使用了两个双向 LSTM 网络分别对源语言句子，机器翻译译文句子进行语义表示，得到两者的语义向量，通过向量间的余弦距离作为评价得分的依据。

如图 1 所示，两个 LSTM 网络的结构，网络参数设置等均相同，仅数据输入不同，本文以源语言的前向 LSTM 为例（其他部分的计算方法同理可得）对该模型的实现过程进行叙述。

令 X 和 Y 分别是长度为 T_x 的源语言句子和长度为 T_y 的机器翻译译文句子，即：$X=(x_1,x_2,...,x_{T_x})$、$Y=(y_1,y_2,...,y_{T_y})$，其计算过程如公式（1）到（12）所示。

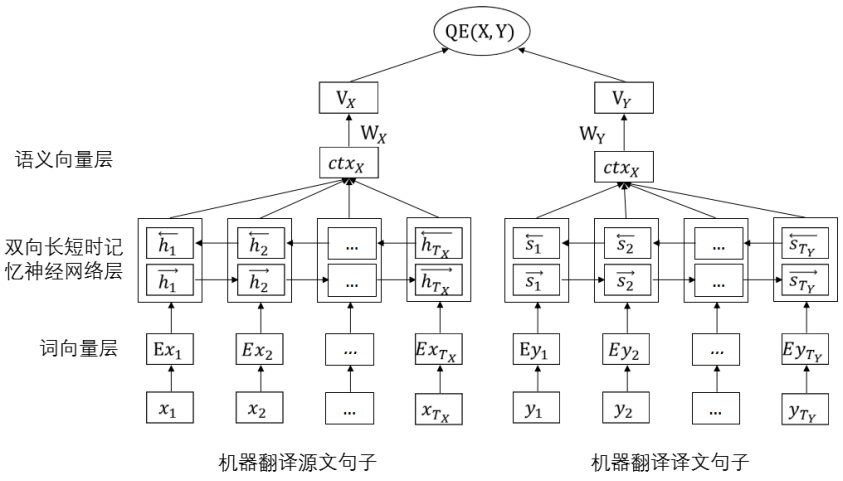

图 1 基于双向 LSTM 的机器翻译自动评价模型结构图

Fig.1 Framework of Quality Estimation with neural network

$$\vec{f}_t=\sigma(\vec{W}_f E_x x_t+\vec{U}_f \vec{h}_{t-1}+\vec{b}_f) \quad (1)$$

$$\vec{i}_t=\sigma(\vec{W}_i E_x x_t+\vec{U}_i \vec{h}_{t-1}+\vec{b}_i) \quad (2)$$

$$\tilde{\vec{C}}_t=\tanh(\vec{W}_C E_x x_t+\vec{U}_c \vec{h}_{t-1}+\vec{b}_C) \quad (3)$$

$$\vec{C}_t = \vec{f}_t * \vec{C}_{t-1} + \vec{i}_t * \tilde{\vec{C}}_t \quad (4)$$

$$\vec{o}_t = \sigma(\vec{W}_o E_x x_t + \vec{U}_o \vec{h}_{t-1} + \vec{V}_o \vec{C}_t + \vec{b}_o) \quad (5)$$

$$\vec{h}_t = \vec{o}_t * \tanh(\vec{C}_t) \quad (6)$$

在公式（1）到（6）中，$\vec{W}_f$、$\vec{U}_f$、$\vec{W}_i$、$\vec{U}_i$、$\vec{W}_C$、$\vec{U}_C$、$\vec{W}_o$、$\vec{U}_o$、$\vec{V}_o$ 均为 LSTM 中的网络权重，$\vec{b}_f$、$\vec{b}_C$、$\vec{b}_o$ 均为网络中的偏置，$\vec{f}_t$ 是网络中忘记门的值，$\vec{i}_t$ 是网络中输入门的值，$\vec{o}_t$ 是网络中输出门的值，$\tilde{\vec{C}}_t$ 是网络记忆单元候选状态，$\vec{C}_t$ 是网络记忆单元更新后的状态，$\vec{h}_t$ 表示 t 时刻前向 LSTM 网络的隐层状态，x_t 表示源语言端网络 t 时刻网络的输入，E_x 表示源语言端词向量矩阵。

最终，双向 LSTM 层得到的输出 h_t 定义成公式（7）所示，在公式（7）中，$\vec{h}_t$ 与 $\overleftarrow{h}_t$ 分别表示前向与后向 LSTM 神经网络在 t 时刻的隐层状态。

$$h_t = [\frac{\vec{h}_t}{\overleftarrow{h}_t}] \quad (7)$$

在语义向量层，将双向 LSTM 得到的隐层状态做平均，接着通过训练两个权重矩阵将源语言与机器翻译译文的表示转换到同一个向量空间表示，计算方法如公式所示。

$$ctx_x = \frac{1}{T_x}\sum_{l=1}^{T_x}(h_{tl}) \quad (8)$$

$$ctx_y = \frac{1}{T_y}\sum_{l=1}^{T_y}(h_{tl}) \quad (9)$$

$$V_x = \sigma(W_x * ctx_x) \quad (10)$$

$$V_y = \sigma(W_y * ctx_y) \quad (11)$$

在公式（8）到（11）中，ctx_x 与 ctx_y 分别表示源语言句子与机器翻译译文的语义向量，V_x 与 V_y 表示在同一个向量空间下的源语言句子与机器翻译译文的向量表示，其中 W_x 与 W_y 为在模型训练过程中需要学习的权重矩阵。

基于以上得到的语义向量表示，定义机器翻译评价得分如公式（12）所示，

$$QE(X,Y) = 1 - \cos(V_x, V_y) \quad (12)$$

在公式（12）中，$\cos(V_x, V_y)$ 定义成 V_x 与 V_y 的余弦相似性，即为两个向量的余弦距离，$QE(X,Y)$ 定义成句对<X,Y>的翻译评价得分。

2 面向目标数据集的伪数据构造

2.1 伪数据构造的基本思想

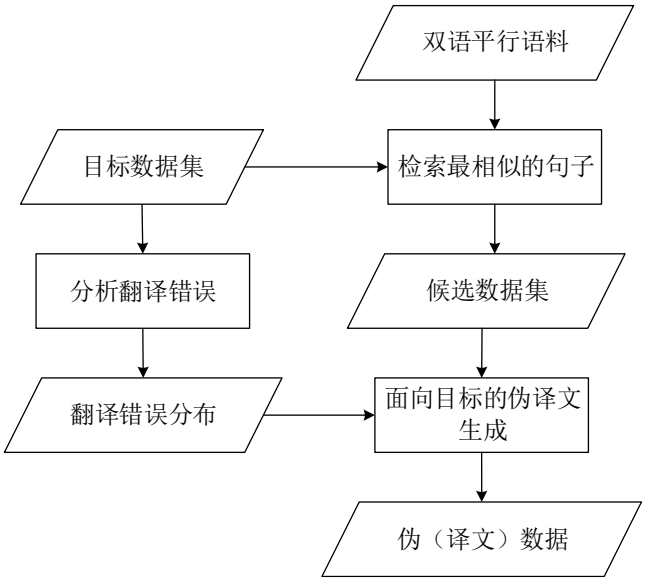

图 2 伪数据生成过程示意图

Fig.2 Sketch map of pseudo data generation

为使得神经网络能学习到各种可能翻译错误及其对应的得分，训练数据应由大量具有不同得分的句对（从源语言到目标语言）组成，而通过真实的机器翻译系统得到大量的机器译文及相应得分通常需要大量的时间。对此，本文通过对平行语料中的句对进行简单的错误编辑操作，从而快速的获得大量的具有不同得分的训练语料。

本文提出了一种面向目标数据集的伪数据生成方法，即根据目标数据集的翻译错误情况构造得到具有不同得分的伪机器译文。如图 2 所示，伪数据生成的流程如下：

1、分析目标数据集的翻译错误。具体包括：翻译错误的数量（翻译错误包括：插入词、删除词、词替换、词调序），翻译过程中错误词的类型（包括名词、动词、代词等）。

2、从双语平行语料中检索出与目标数据集最相似的 n 个句子以获得候选数据集。

3、依据目标数据集的翻译错误分布情况对候选数据集进行编辑（编辑操作包含插入词、删除词、词替换、词调序），从而得到伪机器译文。

基于以上流程完成伪译文数据构造后，将通过基于伪数据的模型预训练与基于训练集的模型精调对翻译质量估计模型进行训练。

2.2 伪数据的生成

在伪数据的生成过程中，主要由三个关键部分构成：目标数据集与候选数据的选取、伪译文的构造以及伪译文得分的构造。其具体实现过程描述如下：

1、目标数据集与候选数据集的选取

在伪数据的构造过程中，目标数据集的翻译错误情况是伪数据构造的依据。因此，目标数据集的翻译错误情况应具有较好的代表性，即目标数据集应尽可能的覆盖不同的翻译错误。鉴于对目标数据集性能的要求，本文选取了 WMT2015 质量估计任务中提供的开发集数据为依据，该数据集包含机器翻译源文、机器翻译译文、标准参考译文由翻译编辑率（Translation Edit Rate，简称 TER）计算工具计算得到的译文得分。此外，该数据集的机器翻译译文是实际的翻译系统产生的，因此，其翻译错误具有一定的代表性。

为了更好的拟合目标数据集的翻译错误情况，对于候选数据集而言，其句子本身应尽可能与目标数据集相似。因此，本文以目标数据集中的每对双语句子作为检索条件，在双语平行语料 Europarl v7[5]（英语到西班牙语）中分别选取相似度最高前 n 个句子组成候选数据集。本文选取了信息检索领域的开源搜索引擎 Indri[6]工具实现该操作。

2、伪译文的构造

- 目标数据集翻译错误情况统计

本文选取 TER 计算工具 tercom 对目标数据集中从参考译文到机器译文的翻译编辑情况进行统计，根据统计结果得到翻译错误的分布。

- 候选数据集的错误编辑

依据统计得到的目标数据集翻译错误分布情况，本文设计了如表 1 所示的错误编辑方案，以此实现对机器译文的构造。以插入词为例，首先以表 1 中的构造方案确定插入词的数量，插入词的位置等信息，在表 1 中，词占句长比意为目标数据集中错误词数量与句子长度（句子词总数）的比例。接着从词表中选取句子中未出现的词进行插入。删除词、替换词、词调序等操作的实现同理可得。

表 1　伪数据构造方案

Table 1　the construction scheme of pseudo data

操作	数量	类型	位置
插入词	句长*词占句长比	随机选取句中未出现的词	随机选取
删除词	句长*词占句长比	随机选取句中未出现的词	随机选取
替换词	句长*词占句长比	随机选取句中未出现的词	随机选取
词调序	句长*词占句长比	随机生成调序长度	随机选取

3、伪译文得分的构造

在该阶段，需要将候选数据集中的目标语言句子为参考译文，将其经过错误编辑操作后得到的句子作为伪译文，利用 TER 计算工具 tercom 计算从参考译文到机器翻译译文的翻译编辑率，以此作为伪译文得分。

2.3 伪数据构造方法的可靠性分析

与开发集中 HTER 的计算方法相同，本文选取了开源 TER 计算工具 tercom[4]作为计算机器翻译译文的得分的依据。但当多种编辑操作组合时，实际的编辑操作与计算工具得到的结果可能存在一定的差异。因此，在本节中，对多种操作组合时，tercom 计算工具的可能产生的误差进行了分析，在伪数据的生成过程中，将尽量避开该计算工具可能存在误差的相关操作，从而得到更具可靠性的伪数据。

为了分析上述选取的 TER 计算工具的误差，本文采用手工标记句子相关错误的方法，设计了相关实验分析该工具的误差。其分析结果如表 2 所示。表 2 给出了两种编辑操作组合时，TER 计算工具给出的结果与实际结果的差异，并给出了误差产生的原因。

表 2　TER 计算工具误差分析

Table 2. Error analysis results of TER calculation tools in operation combination.

实际操作	检测得到的操作	误差产生的原因
插入+删除	无	新插入的词被删除
插入+替换	插入	新插入的词被替换
插入+调序	插入	新插入的词被调序
删除+替换	删除	新替换的词被删除
删除+调序	删除	新调序的词被删除
替换+调序	替换+调序	无

结合 TER 误差分析结果以及表 1 错误生成方案，当同时需要出现多种错误操作时，错误生成操作顺序是：删除词，替换词，词调序，插入词。

3　伪数据下模型的训练与实验结果分析

3.1 模型训练数据集与评价指标

在伪数据的构造过程中，为研究不同伪数据规模下训练得到模型的性能变化情况，本文通过从平行语料中抽取数据规模不同的句子来实现对不同规模的伪数据构造。

为了评价训练得到的模型，本文使用了使用最为广泛的机器翻译质量估计的评价指标 Spearman 秩向相关系数对模型进行评价。

为了研究构造得到伪数据的性能，本节设置了在不同伪数据规模下对模型进行训练的多组对比实验。其中，在基于伪数据的预训练过程中，设置了从 5 万到 160 万等多个数据规模下的预训练实验。在基于真实训练数据进行精调的过程中，采用的数据集为 WMT2015 提供的质量估计任务提供的训练集，训练数据集的详细信息如表 3 所示。

表 3　模型训练数据集

Table 3　Data Setting

训练过程	数据	数量（句对）	描述
预训练	伪数据	5 万~160 万	源语言，目标语言、伪机器翻译译文、TER 得分
模型精调	WMT2015 质量估计任务的训练集	11,271	源语言，机器翻译译文，参考译文、机器翻译得分
测试	WMT2015 质量估计任务的测试集	1817	源语言，机器翻译译文，参考译文

3.2 模型训练设置

对于模型的训练，本文采取伪数据下的预训练与真实训练数据下的精调相结合的两阶段训练方法。

3.2.1　预训练

对于伪数据中的一组数据<X,Y,TERscore>，其中 X 机器翻译的源语言句子，Y 是构造得到的机器翻译译文句子，TERscore 是译文打分，模型训练的目标是去拟合译文打分，定义损失函数如式（15）所示：

$$\ell(X,Y)=(TERscore-QE(X,Y))^2 \tag{15}$$

在公式（15）中，$\ell(X,Y)$ 表示训练过程中的损失函数值，$QE(X,Y)$ 表示模型的给出的分数。

经过预训练学习得到的神经网络将用于模型精调阶段网络的初始化。

3.2.2　模型精调

在预训练模型得到的网络权重的基础上，本文利用机器翻译质量估计任务集下的训练数据进行模型微调。该阶段将训练数据集中提供的 HTER 得分作为人工评价得分，让模型在训练过程中去拟合人工评价得分。此时，模型的训练数据包括：源语言句子，机器翻译译文句子与 HTER 得分。训练目标是让模型给出的得分去拟合 HTER 得分，因此，定义了损失函数如公式（16）所示：

$$\ell(X,Y)=(HTER(X,Y)-QE(X,Y))^2 \tag{16}$$

在公式（16）中，$\ell(X,Y)$ 表示训练过程中的损失函数值，$HTER(X,Y)$ 表示句对<X,Y>的人工评价得分，$QE(X,Y)$ 表示模型给出的得分。

3.2.3　其他参数设置

在神经网络的训练过程中，本文采用了带 Mini-batch 的随机梯度下降（stochastic gradient descent）算法进行模型的训练，并在自适应学习率算法 adadelta[7]下实现训练的优化。

3.3 实验和结果分析

在实验阶段，本文分别从三个角度进行了实验分析。首先，本文对已有的模型训练方法进行了实验分析，主要包括：Zhu[2]等提出的基于双语随机数据预训练的训练策略，以及直接采用真实机器译文进行模型的训练，其实验结果见表 4 所示。

在该阶段的实验中，对于双语随机数据，本文从双语平行语料 Europarl v7[5]抽取得到；对于真实机器译文，本文首先利用开源神经机器翻译系统 Nematus[10]训练得到了一个从英语到西班牙语的机器翻译系统，以伪数据中的源语言句子作为输入，得到西班牙语的机器译文，并计算其 TER 得分，从而完成机器翻译质量估计模型的训练。

表 4　现有模型训练策略得到的结果

Table 4 Result of model training under existing method

预训练数据	数据规模	模型训练结果（预训练+精调）
机器译文	5 万句对	0.12
机器译文	10 万句对	0.13
机器译文	15 万句对	0.15
机器译文	20 万句对	0.22
双语随机数据	20 万句对	0.23
双语随机数据	190 万句对	0.25

其次，为了研究本文构造得到的伪数据的性能，本文在不同伪数据规模下进行了实验，实验结果见表 5 所示。

表 5　基于伪数据的模型训练结果

Table 5 Model training results based on pseudo data

数据规模	预训练结果	精调结果
5 万句对	0.06	0.17
10 万句对	0.12	0.23
15 万句对	0.07	0.23
20 万句对	0.11	0.28
40 万句对	0.13	0.30
80 万句对	0.12	0.27
160 万句对	0.10	0.26

根据对表 5 的实验结果的分析可知，在 20 万伪数据规模下，即可得到比已有训练策略更好的性能。此外，由表 5 可知，预训练结果与模型精调呈现正相关的关系。此外，在构造伪数据的过程中，并非数据量越大越好，当 n 值越大时，所构造得到的伪数据的噪声将被放大，从而导致模型性能下降。

最后，本文还对不同的伪数据构造方案进行了实验分析，即以分析不同的伪数据构造方法下得到的伪数据性能的差异。在该阶段，本文以表 5 中得到的具有最佳性能的数据规模进行实验分析，即在 40 万伪数据规模下进行实验。首先，本文对单一的编辑操作（包括插入、删除、替换、调序）产生的伪数据的性能进行了实验分析。此外，本文还对不同的错误词数量进行了实验对比，结果见表 7 所示。

表 6 仅包含单一编辑操作的伪数据实验结果

Table 6. Result of training under pseudo data only with one type error.

编辑操作	数量	词类型	实验结果（预训练）
插入词	句长*目标数据集每句的错误词数占句长比	随机选取句中未出现的词	0.1
删除词			0.08
替换词			0.02
词调序			0.07
插入词	句长*目标数据集各类错误词数量的均值	随机选取句中未出现的词	0.02
删除词			0.03
替换词			0.02
词调序			0.03

根据表 6 的实验结果可知，仅拟合目标数据集句子单一编辑操作得到的伪数据，其性能差于拟合目标数据集句子中所有编辑操作得到的伪数据。此外，在构造伪数据时，错误编辑词数量的数量去拟合目标数

据集中每句的错误词数占句长的比所得到的伪数据具有更好的性能。

4 结语

为向神经机器翻译质量估计模型提供高效的训练数据，本文提出了面向目标数据集的伪数据构造方法，完成了不同数据规模下的伪数据的构造，通过基于伪数据预训练与模型精调结合的模型训练方法，对机器翻译质量估计模型进行了实验与分析，实验表明，构造得到的伪数据具有高效的性能。在构造得到的伪数据下训练得到的模型，取得了较好的性能。

参考文献

[1] Sutskever I, Vinyals O, Le Q V. Sequence to sequence learning with neural networks[C]//Advances in neural information processing systems. 2014: 3104-3112

[2] Zhu J, Yang M, Li S, et al. Learning Bilingual Sentence Representations for Quality Estimation of Machine Translation[C]//China Workshop on Machine Translation. Springer, Singapore, 2016: 35-42

[3] Kim H, Lee J H. A Recurrent Neural Networks Approach for Estimating the Quality of Machine Translation Output[C]// Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2016

[4] Kreutzer J, Schamoni S, Riezler S. Quality estimation from scratch (quetch): Deep learning for word-level translation quality estimation[C]//Proceedings of the Tenth Workshop on Statistical Machine Translation. 2015: 316-322

[5] Langlois, D.: LORIA System for the WMT15 Quality Estimation Shared Task.In: Proceedings of the Tenth Workshop on Statistical Machine Translation. pp.323{329. Association for Computational Linguistics, Lisbon, Portugal (sep 2015)

[6] Snover M, Dorr B, Schwartz R, et al. A study of translation edit rate with targeted human annotation[J]. Machine Translation Workshop North Bethesda Md, 2006(1):223--231

[7] Koehn P. A parallel corpus for statistical machine translation[J]. Proceedings of the Third Workshop on Statistical Machine Translation, 2005(1):3-4

[8] Strohman T, Metzler D, Turtle H, et al. Indri: A language model-based search engine for complex queries[C]//Proceedings of the International Conference on Intelligent Analysis. 2005, 2(6): 2-6

[9] Zeiler, M.D.: Adadelta: an adaptive learning rate method. arXiv preprint arXiv:1212.5701 (2012)

[10] Sennrich R, Firat O, Cho K, et al. Nematus: a toolkit for neural machine translation[J]. arXiv preprint arXiv:1703.04357, 2017