北京航空航天大学 CCMT2020 质量评测

张文超，巢文涵*，黄彦

（北京航空航天大学 自然语言处理实验室，北京 100191）

摘要：本文介绍了我们参加 CCMT2020 句子级翻译质量评估任务所提交的系统。而在此次评测任务中，本研究组所使用的系统是基于预测器-估计器[1]的体系结构，该结构主要是复现了 CWMT2019 质量评估任务中小牛翻译的模型框架[2]。对于预测器，采用深层 Transformer[3]以及 Transformer-DLCL[3]（先前层的动态线性组合）作为特征提取模型。并且使用从左到右和从右到左的两个翻译模型来获得双向翻译的信息。对于估计器，使用 2 层双向 GRU 来预测句子级任务的 HTER 分数或单词级任务的 OK / BAD 标签。我们先用大规模双语数据对预测器进行预训练，然后将预测器和估计器与 QE 任务数据一起进行联合训练。在本文的其余部分介绍了本组参加评测任务的系统框架、处理方法和评测结果。

关键词：机器翻译；质量评估；深层神经网络

中图分类号：TP391　　**文献标识码：**A

1 引言

质量评估（QE）的目标是在无需访问参考翻译译文的情况下来评估机器翻译系统的输出，在指导提高机翻系统性能和进行后期编辑方面发挥着重要的作用。到如今，质量评估(QE)的发展可以大致分为三个阶段：基于传统机器学习的方法、基于传统机器学习和神经网络相结合的方法、基于纯粹神经网络的方法。

传统的机器学习方法主要的研究内容在于手工特征的选择和抽取。其研究的QE特征包括基线特征（Specia 等人 2013）[4]，语言特征（Felice 和 Specia 2012）[5]，主题模型特征（Rubino 等人 2013）[6]和伪参考特征（Soricut 和 Echihabi 2010; Kozlova et al.2016）[7-8]。多项研究还通过应用特征选择方法来利用潜在特征：主成分分析（González-Rubio 等人，2012）[9]，高斯过程（Shah 等人，2015）[10]和偏最小二乘回归（González-Rubio 等人，2013）[11]。而一个典型的框架则是 QUEST ++ [12]，它提供了多种功能和机器学习方法来构建 QE 模型。

而将QE作为回归/分类任务时会出现两个问题，第一个问题是 QE 数据的可用大小仍然很小，QE 数据昂贵且不易获得。第二个问题是，大多数传统的质量估计方法都是基于浅层架构的，并且依赖手工设计的特性来捕获特性集和 QE 注释之间的复杂关系[13]。而最近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难题，很多研究者尝试使用深度神经网络自动抽取质量评估特征并完成评分。代表性的工作包括 Kreutzer 等人在 2015 年为 QE 提出了一种基于窗口的 FNN 体系结构，称为 scraTCH 的质量估计（QUETCH[14]），Patel 和 M（2016）提出了一种基于 RNN 的 QE 架构[15]，Martins 等人（2016）提出了 QUETCH 的三个扩展[16]。在该阶段中，为了获得双语上下文窗口，还需要 SMT 中的单词对齐组件，从而使生成的模型“不完全”成为神经网络模型。

在基于 Predictor-Estimator[1]的神经网络方法中，词预测模型和质量估计模型均由神经网络组成。该结构的主要组成部分是一个基于改进的神经机器翻译模型的词预测模型，这是一个基于不需要双语单词对齐组件的双向双语递归神经网络 RNN 语言模型，可以根据源词和目标词的上下文来预测目标词。所提出的质量估计方法依次训练以下两类神经模型：（1）预测器：由平行语料库训练的神经词预测模型；（2）估计器：由质量估计数据训练的神经质量估计模型。为了将预测任务转化为质量估计任务，先从质量估计模型中生成质量估计特征向量，再将其输入到质量估计模型中。

在我们的工作中，我们提交的任务也是基于 Predictor-Estimator 的模型架构。我们使用大量的双语数据来预训练从左到右和从右到左的深层Transformer词预测模型。之后，将多层 Bi-GRU 用作估计器，并使用 QE 数据与预测器进行联合训练来完成单词

通信作者：chaowenhan@buaa.edu.cn

级的分类任务和句子级的回归任务。

2 Predictor-Estimator 系统

Predictor-Estimator 是一个两阶段的端到端神经网络模型，该模型包含两个子模型：在额外大规模平行语料上进行训练的单词预测模型（Predictor），在平行 QE 数据上进行训练的质量评估模型（Estimator）。Predictor 是基于传统 Transformer 模型改进后的深层机翻模型，包括 12r 和 r21，通过根据源句和目标句子的上下文来预测目标句子中的每一个单词，得到 QE 特征向量(QEFVs)，然后 QEFVs 作为 Estimator 的输入来估计翻译质量。该模型的网络结构如图 1 所示。

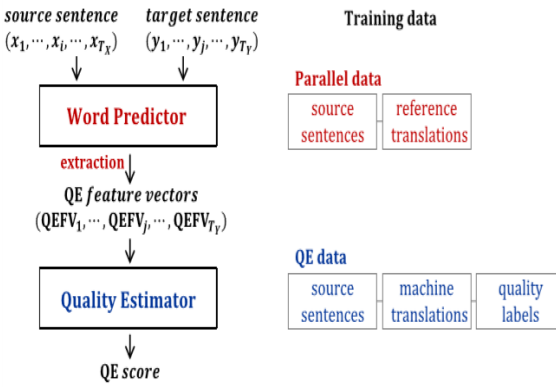

图 1 Predictor-Estimator 结构图以及两阶段的数据使用情况

定义源语言句子为 X=（x1，...,xi,...,xTx），目标语言句子为 Y=（y1,...,yj,...,yTy），其中 xi 是源语言句子中的第 i 个单词，yj 是目标语言句子中的第 j 个单词，Tx 与 Ty 分别是源语言句子和目标语言句子的长度。

2.1 深层 Transformer

该模型区别于 Transformer-Base 和 Transformer-Big 模型，其采用更深层的 Transformer-DLCL（先前层的动态线性组合）作为特征提取模型，其中 encoder 层为 25~30 层。对于 Transformer，学习较深的网络并不容易，因为由于梯度消失/探索问题而难以优化。但是 Wang[3]等人强调层规范化的位置在训练深层 Transformer 时起着至关重要的作用。实验发现将层归一化应用于每个子层的输入，这可以提供一种直接方法来从上到下传递误差梯度。这样，当模型更深入时，规范前的 Transformer 比规范后的（vanilla Transformer）更有效地进行训练。

另外，在 Transformer 模型中使用了先前各层的动态线性组合方法，Transformer-DLCL 采用与所有先前层的直接连接，并提供对深层堆栈中较低层表示的有效访问。使用附加的权重矩阵$W_{l+1} \in R^{L\times L}$以线性方式加权每个传入层。

为了获得双向翻译的信息以及充分提取上下文知识，效仿阿里团队 Bilingual Expert[17]系统使用从右到左和从左到右的两个翻译模型，并分别提取特征向量 12r 和 r21。 最后通过级联的方式得到最终的质量向量（q = [12r：r21]）。

2.2 提取 QE 特征向量：QEFVs

对于提取 QEFVs，需要指出的是在 Predictor 阶段用的平行语料和在 Estimator 阶段用的 QE 数据是存在差异的：平行语料中只有正确的句子（源文、参考译文），而 QE 数据中存在正确的和不正确的句子（源文、机翻译文）。基于此差异，(Kim et al.,2017[1]) 提出两种类型 QEFVs：预测前 QEFVs（Pre-QEFVs）和预测后 QEFVs（Post-QEFVs）。

对于目标语句中第 j 个位置的单词，有

$$\text{Pre-QEFVj} = W_{yj} * fj \quad (1)$$

$$\text{Post-QEFVj} = [\overrightarrow{sj}\ ;\overleftarrow{sj}\ ] \quad (2)$$

其中 W_{yj} 是对应 yj 的特征权重，fj 是表示 yj 的特征向量，$\overrightarrow{sj}$和$\overleftarrow{sj}$是 yj 的前向和后向隐藏层状态。

2.3 Bi-GRU Estimator

因为 QE 数据和平行语料在质量方面是有差异的，所以需要将得到的 QEFVs 作为 Estimator 阶段的输入向量进行训练，实现在 QE 数据上的模型微调。

Estimator 是一个基于 Bi-GRU 的网络模型，将句子 Y 的 QEFV 序列

QEFVs(QEFV1,...,QEFVTy)作为输入，得到QEFVs的隐藏层表示，然后在不同粒度的评估任务中对该隐藏层表示做不同的处理以完成质量评估。

Estimator阶段的网络结构如图2所示：

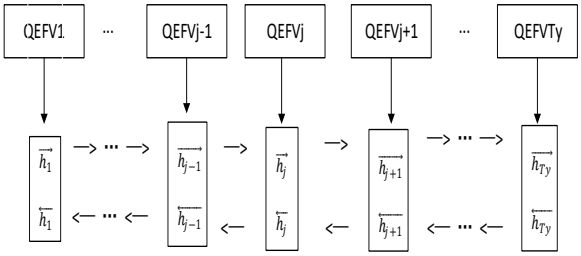

图2　Estimator网络结构

2.3.1 句子级质量评估

因为句子级质量评估的目标是预测目标语言译文的HTER值（在0到1之间），所以在该阶段应用Bi-GRU网络将QEFVs(QEFV1,...,QEFVTy)先转换成一个单一向量S，然后将其看作是逻辑回归任务。

其中向量S是由最后两个隐藏层状态$\overrightarrow{h_{Ty}}$和$\overleftarrow{h_1}$拼接得到的。

句子级质量评估模型定义如下：

$$QE_{sentence}(Y, X) = QE_{sentence}(QEFV_1, \ldots, QEFV_{Ty}) = \sigma(Ws * S) \quad (3)$$

其中Ws是权重矩阵。

2.3.2 单词级质量评估

单词级质量评估的目标是给每个token赋予二分类标签OK/BAD。不同于句子级质量评估的是，在得到目标语言Y第j个位置的隐藏层状态h_j之后不会进行拼接，而是直接作为该位置的S向量进行计算。第j个单词位置的质量评估模型定义如下：

$$QE_{word_j}(Y, X) = QE_{word_j}(QEFV_1, \ldots, QEFV_{Ty}) = \begin{cases} OK & \text{如果 } \sigma(Ww * h_j) \geq \textit{阈值} \\ BAD & \text{如果 } \sigma(Ww * h_j) < \textit{阈值} \end{cases} \quad (4)$$

其中Ww是权重矩阵，阈值定为0.5。

3 实验与分析

3．1 实验数据

对于预训练数据，我们使用近两年CWMT EN-ZH机翻评测任务平行数据以及自行搜集的UNcorpus对我们的预测器进行预训练。筛选后，选择了大约12M个句子对。所有平行数据都通过官方给出的单词分段工具包进行分段，预处理后训练BPE[18]模型。

QE任务的数据集包括三个部分：源句子，机器翻译和QEscore（句子级别的HTER得分或单词级别的OK / BAD标签）。 CCMT 2020 QE任务提供的数据量不超过15K，因此我们把CWMT 2019中的QE数据也加进来进行训练。

3．2 模型参数

我们基于Fairseq[19]实施我们的QE模型。 预训练模型在两块TITAN Xp上进行了训练。我们使用Adam优化器，其中β1= 0.97，β2= 0 .997和ε = 10^{-6}。我们将最大学习率设置为0.002，将warmup-steps设置为8000。对于联合训练预测器-估计器体系结构，我们将其训练在一个TITAN Xp GPU上，我们将最大学习率设置为 0.0005，并将warmup-steps设置为200。

3．3 质量评估指标

为了进行评估，我们使用了CCMT QE共享任务的官方评估措施，定义如下：

3.3.1 句子级别评估措施

Pearson相关性：目标翻译的预测HTER分数与真实HTER分数（金标签）之间的线性相关性的度量。

3.3.2 单词级别评估措施

（1）F1-BAD和F1-OK的乘法（F1-mult）：乘法F1-BAD和F1-OK，它们有两个相互平衡的组件。

（2）F1-BAD：“BAD”标签的F1分数。

（3）F1-OK：“OK”标签的F1分数。

3．4 实验结果

英中dev验证集质量评估结果如表1所示。

表1 英中dev验证集质量评估结果

英中句子级别	Pearson's
Sent	0.54769
中英句子级别	Pearson's
Sent	0.57312

3.5 分析

从上面表中可以看出，本研究组的实验结果并不是十分理想，经分析认为影响因素有以下几点：首先，在 Predictor 阶段，所使用的预训练数据没有经过严格的筛选，所以提取的质量特征不够精确。其次，没有对模型进行 ensemble 操作。在使用 QE 数据时没有精准分析数据的分布特性。

4 结论

本文介绍了我们用于 CCMT2020 质量评估任务的系统。我们采用 predictor-estimator 体系结构，使用基于深度网络的 Pre-norm Transformer-DLCL 作为 Predictor，并结合了 L2R 和 R2L 的模型来进一步增强预测器的特征提取功能。使用 Bi-GRU 作为估计器,并使用预测器提取的质量向量来预测不同的任务。

参考文献

[1] Hyun Kim, Hun-Young Jung, Hong-Seok Kwon,et al.: Predictor-Estimator: Neural Quality Estimation Based on Target Word Prediction for Machine Translation[J]. ACM Trans. Asian & Low-Resource Lang. Inf. Process, 2017, 17(1): 3:1-3:22.

[2] Ziyang Wang, Hui Liu, Hexuan Chen, Kai Feng, Zeyang Wang, Bei Li, Chen Xu, Tong Xiao, and Jingbo Zhu: NiuTrans Submission for CCMT19 Quality Estimation Task.

[3] Qiang Wang, Bei Li, Tong Xiao, Jingbo Zhu, Changliang Li, Derek F. Wong, Lidia S. Chao:Learning Deep Transformer Models for Machine Translation. ACL (1) 2019: 1810-1822

[4] Specia, L., Shah, K., de Souza, J.G.C., Cohn, T., Kessler, F.B.: Quest a translation quality estimation framework. In: Proceedings of the 51st ACL: System Demonstrations, pp. 79-84 (2013)

[5] Mariano Felice and Lucia Specia. 2012. Linguistic features for quality estimation. In Proceedings of the 7th Workshop on Statistical Machine Translation. Association for Computational Linguistics, 96-103.

[6] Raphael Rubino, Jose de Souza, Jennifer Foster, and Lucia Specia. 2013. Topic models for translation quality estimation for gisting purposes. In Proceedings of the XIV Machine Translation Summit. 295-302.

[7] Radu Soricut and Abdessamad Echihabi. 2010. TrustRank: Inducing trust in automatic translations via ranking. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 612-621.

[8] Anna Kozlova, Mariya Shmatova, and Anton Frolov. 2016. YSDA participation in the WMT'16 quality estimation shared task. In Proceedings of the 1st Conference on Machine Translation. Association for Computational Linguistics, 793-799.

[9] Jesús González-Rubio, Alberto Sanchís, and Francisco Casacuberta. 2012. PRHLT submission to the WMT12 quality estimation task. In Proceedings of the 7th Workshop on Statistical Machine Translation. Association for Computational Linguistics, 104-108.

[10] Kashif Shah, Trevor Cohn, and Lucia Specia. 2015. A bayesian non-linear method for feature selection in machine translation quality estimation. Mach. Transl. 29, 2 (2015), 101-125. DOI:http://dx.doi.org/10.1007/s10590-014-9164-x

[11] Jesús González-Rubio, J. Ramón Navarro-Cerdán, and Francisco Casacuberta. 2013. Dimensionality reduction methods for machine translation quality estimation. Mach. Transl. 27, 3 (2013), 281-301. DOI:http://dx.doi.org/10.1007/s10590-013-9139-3

[12] Lucia Specia, Gustavo Paetzold, Carolina Scarton: Multi-level Translation Quality Prediction with QuEst++. ACL (System Demonstrations) 2015: 115-120

[13] Hyun Kim, Hun-Young Jung, Hong-Seok Kwon, Jong-Hyeok Lee, Seung-Hoon Na: Predictor-Estimator: Neural Quality Estimation Based on Target Word Prediction for Machine Translation. ACM Trans. Asian & Low-Resource Lang. Inf. Process.17(1): 3:1-3:22 (2017)

[14] Julia Kreutzer, Shigehiko Schamoni, and Stefan Riezler. 2015. QUality estimation from ScraTCH (QUETCH): Deep learning for word-level translation quality estimation. In Proceedings of the 10th Workshop on Statistical Machine Translation. Association for Computational Linguistics, 316-322.

[15] Raj Nath Patel and Sasikumar M. 2016. Translation quality estimation using recurrent neural network. In Proceedings of the 1st Conference on Machine Translation. Association for Computational Linguistics, 819-824.

[16] André F. T.Martins, Ramón Astudillo, Chris Hokamp, and Fabio Kepler. 2016. Unbabel's participation in the WMT16 wordlevel translation quality estimation shared task. In Proceedings of the 1st Conference on Machine Translation. Association for Computational Linguistics, 806-811.

[17] Kai Fan, Bo Li, Fengming Zhou, Jiayi Wang:Bilingual Expert Can Find Translation Errors. CoRR abs/1807.09433 (2018)

[18] Sennrich, R., Haddow, B., Birch, A.: Neural machine translation of rare words with subword units. arXiv preprint arXiv:1508.07909 (2015)

[19] Ott, M., et al.: fairseq: A fast, extensible toolkit for sequence modeling. In: Proceedings of NAACL-HLT 2019: Demonstrations (2019)

Beihang University CCMT2020 Evaluation Technology Report of Translation Quality

ZHANG Wenchao, CHAO Wenhan*, HUANG Yan

(Beihang University, Natural Language Processing Labs, Beijing, 100191,China)

Abstract: This article introduces the system we submitted for the CCMT2020 sentence-level translation Quality Estimation Task. In this evaluation task, the system used by our group is based on the Predictor-Estimator[1] architecture, which mainly reproduces the model framework of NiuTrans translation in the CWMT2019 quality estimation task[2]. For the Predictor, Deep Transformer[3] and Transformer-DLCL[3] (dynamic linear combination of previous layers) are used as feature extraction models. And use the left-to-right and right-to-left two models to obtain bidirectional translation information. For the Estimator, a 2-layer bidirectional GRU is used to predict the HTER score of sentence-level tasks or the OK/BAD label of word-level tasks. We first use large-scale bilingual data to pre-train the Predictor, and then jointly train the Predictor and Estimator with the QE task data. In the rest of this article, we will introduce the system framework, processing methods and evaluation results in the evaluation task.

Key words: Machine Translation；Quality Estimation；Deep Neural Network