文章编号：2095-2163(2019)02-0271-05 中图分类号：TP391.4 文献标志码：A

融合翻译知识的机器翻译质量估计算法

孙 潇，朱聪慧，赵铁军
（哈尔滨工业大学 计算机科学与技术学院，哈尔滨 150001）

摘 要：句子级别的机器翻译质量估计（Sentence-Level Translation Quality Estimation）任务以源语言句子及相应机器翻译系统的结果为输入，对译文的质量进行估计。针对一些使用神经网络和词向量的基于特征工程的模型中，词向量不包含机器翻译知识的问题，本文提出了一种融合机器翻译知识的翻译质量估计方法。该方法将双向 NMT 模型融入质量估计方法的训练过程中。实验结果表明，翻译质量估计的 Pearson 相关系数有所提升，证明了所提方法的有效性。

关键词：翻译质量估计；神经网络机器翻译模型；双向 NMT 模型

Translation quality estimation method that combines machine translation knowledge

SUN Xiao, ZHU Conghui, ZHAO Tiejun
(School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China)

【Abstract】 Sentence-level machine translation quality estimation (Sentence-Level QE) takes the sentence of source language and the corresponding machine translation as input, and estimates the quality of the translation. Aiming at the problem that some feature engineering based methods using neural network and word embedding don´t include machine translation knowledge, this paper proposes a translation quality estimation method that combines machine translation knowledge.The method incorporates bidirectional neural machine translation models into the training process of the translation quality estimation model. The experimental results show that the Pearson correlation coefficient for HTER prediction has been improved, which proves the effectiveness of the proposed method.

【Key words】 translation quality estimation; neural network machine translation model; bidirectional NMT models

0 引 言

随着经济的发展，国际交流合作日益频繁，对机器翻译的需求逐渐增大。而机器翻译译文质量的自动评价，对机器翻译的研究非常重要。其中，广泛使用的 *BLEU* 评价指标就推动了机器翻译的进步与发展。

目前常用的 *BLEU* 评价指标存在 2 个主要问题。首先是指标的计算要求有参考译文作为输入，其次指标在句子级别上对译文的评分效果比较差。而句子级别的机器翻译质量估计（Sentence-Level Translation Quality Estimation，Sentence-Level QE）则可显著改善这类现象。Sentence-Level QE 是指在没有参考译文的情况下，只根据源语句，来对机器翻译译文的质量进行估计。定义中的质量可以指：*adequate*（和源语句的意思相近程度）、*fluency*（翻译的流畅程度）、*HTER*（Human-targeted Translation Edit Rate）等等。其中，*HTER* 最为常用。*HTER* 是机器翻译的译文和人工修改的参考译文（Human-targeted Translation）之间的编辑距离除以所有参考译文的平均长度。

以往的基于特征工程的翻译质量估计方法的研究中，一些用神经网络提取特征的方法并没有考虑引入翻译知识。

本文中，研究提出一种原创的用神经网络机器翻译模型来为 QE 任务提取特征的方法，该方法利用了 NMT 模型，比以往的用神经网络提取的 QE 特征包含了更多的语义信息。

1 相关工作

对句子级别的机器翻译质量估计的研究，一般是将其归作为一个有监督的回归问题，此前的研究主要是应用传统的统计模型，比如 SVR、线性回归模型等等，研究均重点致力于特征提取（feature extraction）和特征选择（feature selection）方面。其中，特征提取指的是从源语句和对应的机器翻译的

作者简介：孙 潇（1994-），男，硕士研究生，主要研究方向：自然语言处理、机器学习；朱聪慧（1979-），男，博士，讲师，硕士生导师，主要研究方向：自然语言处理、机器学习；赵铁军（1962-），男，博士，教授，博士生导师，主要研究方向：自然语言处理、机器学习。

收稿日期：2018-06-07

译文以及一些外部的资源或工具中提取构造和译文质量相关的特征,也就是针对这个机器学习任务做特征工程(feature engineering)。而特征选择是指,从已经提取的特征集合中选择预测效果最好的特征子集,这可以看作是一个搜索寻优问题,并被证明是一个 NP 问题,无法在多项式的时间复杂度内得到准确解。因此机器译文质量估计的特征选择一般包括产生候选子集和对特征子集进行评价这 2 个要素,机器译文质量估计领域常用的特征选择算法包括高斯过程[1]、启发式[2]。在之前句子级别机器译文质量估计的研究中,至关重要的即是特征提取,也就是人工设计合适的特征[3-6]。常见的人工提取的特征包括源语句长度、目标语句长度、特殊字符匹配率等等。这些人工提取的特征,大多数是一些语法特征,很少涉及到语句的深层次语义信息。

随着深度学习的发展,有些研究者将神经网络用于特征提取的过程中,然后将提取到的特征单独或者和其它传统特征一同输入到机器学习模型中;常见的神经网络提取的特征包括源语句和目标语句在神经网络语言模型中的分数、在神经网络机器翻译下的分数、语句的所有单词对应的词向量的平均值等等[7-10]。这些特征和之前传统的特征相比,包含了较多的语义信息。

除了用神经网络提取特征,然后应用传统的统计模型外,有的研究更进一步提出了基于多层神经网络的端到端的机器译文质量估计模型[11-14]。而且,研究中 QE 任务的数据集比较小,因此直接训练端到端的模型,将存在过拟合的风险。目前,效果较好的此类方法,一般都是直接或间接地利用了大量的平行语料来提高模型的泛化能力。

2 模型详述

2.1 基本模型简述

本文利用神经网络机器翻译模型来为机器翻译译文质量估计问题(QE)提取特征,是对直接将语句的单词词向量的平均作为特征的方法的有效改进。在本文第一节中提到,QE 领域的研究中,对特征的提取非常关键;在特征提取方面,之前的研究主要是对源语句和机器翻译的译文提取语法相关的特征,也有一些研究探讨了语义问题。随着近些年深度学习的兴起,一些研究使用神经网络来提取和句子的语义相关的特征。其中一个方法是,用词袋模型对句子建立模型,也就是将句子看成是单词的集合,不考虑词语间的先后顺序,用该语句的所有单词对应的词向量的平均值作为对该语句的编码。对源语句和译文用上述方法编码之后,得到 2 个向量,对这 2 个向量进行拼接,作为 QE 模型的输入特征。

这种直接对句子中的单词的词向量求平均的方法,没有考虑词语间的先后顺序和联系,很难提取到语句深层次的语义信息。因此可以考虑用循环神经网络(Recurrent Neural Network, RNN)对句子进行编码,本文采用的是 GRU(Gated Recurrent Unit)。GRU 是循环神经网络的一种,不仅可以适用于如自然语言语句这种变长的序列研究,同时也可以如长短期记忆网络(Long Short-Term Memory, LSTM)一样处理较长距离的依赖关系,但与 LSTM 相比结构更加简单,因此本文在循环神经网络的变体中选用 GRU 作为编码器(和解码器)。同时,针对已有研究的分析表明,GRU 每一步的隐状态包含了输入序列中当前输入以及之前所有输入的信息,因此本文采用 GRU 最后一步输出的隐状态作为对整个语句的编码向量。

此外,因为 QE 任务的数据集一般比较小,比如本文实验选用的训练集只有 2 万个标注数据;而机器翻译领域的常见语言对的数据集一般比较大,因此本文考虑通过引入 2 个简单的神经网络机器翻译(Neural Machine Translation, NMT)模型,来充分利用大量的平行语料。引入的 2 个 NMT 模型翻译方向相反,一个是源端到目标端语言,另一个是目标端语言到源端语言。这 2 个 NMT 模型的编码器分别对源语句和目标语句进行编码得到编码向量,然后 2 个 NMT 模型的解码器再分别对编码向量解码得到目标语句和源语句;其中,2 个 NMT 模型对源语句和目标端语句编码得到的编码向量理论上就分别包含了源语句和目标语句的信息。本文利用 2 个 NMT 模型的编码器分别对源语句和机器翻译的译文进行编码,得到的向量就作为 QE 模型的输入特征。

整个模型由 2 部分构成。第一部分是 2 个翻译方向相反的 NMT 模型,第二部分是 QE 模型,输出最终的质量 *HTER*。输入的是从源语句和目标语句提取得到的特征向量,在这里是 2 个 NMT 模型编码得到的编码向量,特征向量中除此之外也可以包含通过其它途径提取到的特征。整体的模型结构如图 1 所示。

2.1.1 NMT 子模型

整个算法中一共包括 2 个翻译方向相反的 NMT 模型,分别是源端到目标端和目标端到源端。

2 个 NMT 模型结构完全相同，共享词向量参数。下面即以源端到目标端的 NMT 模型为例展开论述。源端的语句 $X = \{x_1, x_2, \cdots, x_S\}$，$x_i(1 \leqslant i \leqslant S)$ 是源语句中的单词的 one-hot 编码，S 为源端语句的长度；目标端语句 $Y = \{y_1, y_2, \cdots, y_T\}$，$y_j(1 \leqslant j \leqslant T)$ 是目标语句中的单词的 one - hot 编码，T 为目标端语句的长度。源端和目标端的词向量矩阵为 $\boldsymbol{E}_S$ 和 $\boldsymbol{E}_T$，其中词向量矩阵的每一列代表一个单词的词向量。选用的 NMT 模型由编码器和解码器 2 部分组成，编码器和解码器使用的神经网络模型都是 GRU。编码器的功能是将源端语句 X 编码为固定向量 $\boldsymbol{C}$。然后解码器对 $\boldsymbol{C}$ 进行解码得到目标端语句 Y。整个 NMT 模型可以表示为 $P(Y \mid X;\theta)$，该条件概率可以用概率的乘法法则分解，数学公式可见如下：

$$P(Y \mid X;\theta) = \prod_{j=1}^{T} p(y_j \mid x, y_1, \cdots, y_{j-1};\theta), \quad (1)$$

其中，编码器主要由 GRU 构成，GRU 初始的隐状态为零向量。在每一步的实际计算中，需先将该步的单词的 one-hot 表示 x_i 用词向量矩阵 $\boldsymbol{E}_S$ 映射为词向量 $\boldsymbol{E}_S \times x_i$，然后和上一步的隐状态一起作为输入，进行 GRU 当前步的计算。并且将最后一步输出的隐状态 h_S 作为对整个源语句的编码向量 $\boldsymbol{C}$。第 t 步的计算公式可表示为：

$$r_t = sigmoid(W_{ir}\boldsymbol{E}_S x_t + b_{ir} + W_{hr} h_{t-1} + b_{hr}), \quad (2)$$

$$z_t = sigmoid(W_{iz}\boldsymbol{E}_S x_t + b_{iz} + W_{hz} h_{t-1} + b_{hz}), \quad (3)$$

$$n_t = tanh(W_{in}\boldsymbol{E}_S x_t + b_{in} + r_t(W_{hn} h_{t-1} + b_{hn})), \quad (4)$$

$$h_t = (1 - z_t) n_t + z_t h_{t-1}, \quad (5)$$

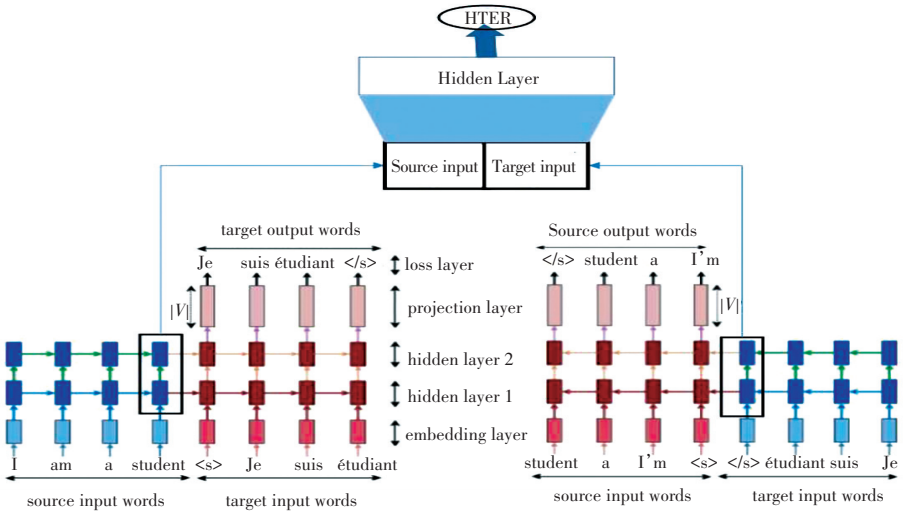

图 1　模型整体结构

Fig. 1　The structure of the model

解码器对源语句的编码向量 $\boldsymbol{C}$ 进行解码。采用的神经网络模型是 GRU，初始的隐状态是 $\boldsymbol{C}$，$\boldsymbol{C}$ 包含了源语句的信息。每一步最终的输出是对这一步的词表中所有单词的概率分布，而输入却是上一步的预测的单词的词向量，训练过程中的输入则是上一步中目标语句对应的单词的词向量。第 t 步的隐状态 h_t 的计算公式和编码器部分相似。这里采用的是一个单隐层的前向神经网络，第 t 步的目标词概率分布的计算公式具体如下：

$$p(y_t \mid y_1, \cdots, y_{t-1}, x;\theta) = softmax(W_{o2}\, tanh(W_{o1} h_t + b_1))_{y_t}, \quad (6)$$

2.1.2　QE 模型

QE 模型的输入是特征向量 $\boldsymbol{V}$，在基本模型中特征向量是源端句子编码向量 $\boldsymbol{C}_S$ 和目标端句子编码向量 $\boldsymbol{C}_T$ 的拼接 $[\boldsymbol{C}_S:\boldsymbol{C}_T]$。模型采用的是单隐层的前向神经网络，权重分别是 W_1 和 W_2，偏置向量分别是 b_1 和 b_2。隐层的激活函数采用 *relu*，输出层因为要输出 0 ~ 1 的分数，因此采用 *sigmoid* 作为激活函数。公式表述如下：

$$a_1 = relu(W_1 \boldsymbol{V} + b_1), \quad (7)$$

$$hter = sigmoid(W_2 a_1 + b_2). \quad (8)$$

2.2　加入其他特征

对源语句和机器翻译译文的编码向量分别包含了源语句和译文的语义语法信息，但是向量的每个维度都具有不可解释性。因此本文将其它一些人工提取的特征和这 2 个用神经网络提取的特征进行连接，作为 QE 模型的输入特征。这些特征都具有高度直观、且容易理解的含义。添加的特征有 17

个[15],对其含义可阐释解析如下。

(1) 源语句中的单词数量。

(2) 机器翻译语句中的单词数量。

(3) 源语句长度。

(4) 源语句的语言模型概率。

(5) 机器翻译语句的语言模型概率。

(6) 机器翻译语句内单词出现次数的平均值。

(7) 源语句中每个单词对应的翻译数量的平均值(使用 IBM 模型 1, 阈值设置为 $prob(t \mid s) > 0.2$)。

(8) 源语句中每个单词对应的翻译数量(使用 IBM 模型 1, 阈值设置为 $prob(t \mid s) > 0.01$) 的加权平均值, 权重为源语言语料库中每个词的逆频率。

(9) 源语句中的单词占源语言语料库(SMT 训练平行语料库) 中频率四分位数 1(频率较低的单词)的百分比。

(10) 源语句中的单词占源语言语料库中频率四分位数 4(频率较高的单词)的百分比。

(11) 源语句中的 bigrams 占源语言语料库中频率四分位数 1 的百分比。

(12) 源语句中的 bigrams 占源语言语料库中频率四分位数 4 的百分比。

(13) 源语句中的 trigrams 占源语言语料库中频率四分位数 1 的百分比。

(14) 源语句中的 trigrams 占源语言语料库中频率四分位数 4 的百分比。

(15)在语料库(SMT 训练平行语料库)中可以看到的源语句中的单词所占的百分比。

(16)源句子中标点符号的数量。

(17)目标语句中标点符号的数量。

3 实验

3.1 实验设置

本文为了对用 NMT 模型提取的特征的效果进行验证,在 2 个不同的数据集上分别进行了 4 组实验,每组实验的不同点主要在于输入的特征。这 4 组实验采用的特征,分别是:17 个人工提取的特征、词向量特征、NMT 模型提取的特征、NMT 提取的特征加上 17 个人工提取的特征。其中,第一组实验采用 SVR 作为模型,其它组的模型采用前向神经网络。这里,关于本次实验中的数值指标设计,对其可概述如下。

(1)模型和训练的参数设置。SVR 的核函数采用径向基,其他超参数使用交叉验证确定。源端和目标端词表大小为 74 000,词向量的维度设置为 512,神经网络(包括 GRU、全连接神经网络)的隐层神经元个数为 1 024。神经网络的优化算法采用 adam,batch 的大小为 64,训练 NMT 模型的学习率为 3e-4,训练 QE 模型的学习率为 5e-5。

(2)实验所使用的数据集描述。用于训练 NMT 的数据集来自于 WMT 2017 shared task 的 en-de 翻译任务,语料包括 Europarl v7、Common Crawl corpus、News Commentary v12、Rapid corpus of EU press releases 等,总共 3 M 个句对。研究采用的 NMT 模型结构比较简单,因此从所有 3 M 个句对中随机抽取 90 w 个句对。再加上对应的 QE 数据集(源语句加上被人工 post edit 后的译文)中的 2 w 个句对,组成训练本文所需的 NMT 模型的平行语料。

用于训练 QE 的数据集来自于 WMT17 Shared Task: Quality Estimation 任务一,包括德语到英语和英语到德语 2 个方向的数据集,并且分别属于 2 个不同的领域。数据集信息详见表 1。

表 1 QE 数据集

Tab. 1 QE data set

	Train	Dev	Test	Domain
QE 数据集(de-en)	23 000	1 000	2 000	Pharmaceutical
QE 数据集(en-de)	25 000	1 000	2 000	IT

3.2 实验结果

实验运行结果参见表 2、表 3。

表 2 de-en 数据集 Pearson 相关系数

Tab. 2 The Pearson of de-en

	Dev	Test
Baseline(17 features)	0.452	0.439
Embedding average	0.501	0.487
NMT-based QE	0.545	0.531
NMT-based QE+17 features	0.561	0.557

表 3 en-de 数据集 Pearson 相关系数

Tab. 3 The Pearson of en-de

	Dev	Test
Baseline(17 features)	0.407	0.401
Embedding average	0.496	0.481
NMT-based QE	0.516	0.502
NMT-based QE+17 features	0.523	0.511

综上结果分析可知,在 2 个方向上,可以看到相比于人工提取的 17 个特征,即使使用词向量直接相加提取的特征,效果也会更好。这说明词向量包含的单词带有大量的语义信息,即使不考虑单词之间

的顺序和关系,也可以对最终译文的质量的预测有所帮助。然后本文使用了 NMT 模型中的编码器对句子的单词序列进行了非线性变换,最终的实验结果表明,这种非线性变换和直接求平均相比,对机器翻译译文质量的预测能力更强。最后,编码器得到的编码向量虽然包含了语义信息,但是每个维度都具有不可解释性,将其和人工提取的 17 个具有直观含义的特征拼接起来作为输入特征,效果有所提升,说明编码向量特征和这 17 个特征在一定程度上实现了互补。

4 结束语

针对机器翻译译文质量估计问题,本文提出了一个融合了翻译知识的特征提取算法,该算法首先训练 2 个翻译方向相反的 NMT 模型,然后利用 2 个编码器编码得到向量作为特征。实验表明,利用 NMT 编码器提取的特征比直接对语句中单词词向量平均的特征预测效果更好。并且,该特征和本文提到的 17 个手工提取的特征一定程度上具有互补性,2 类特征的结合可以进一步提升 QE 模型的效果。

参考文献

[1] SHAH K, COHN T, SPECIA L. A bayesian non-linear method for feature selection in machine translation quality estimation[J]. Machine Translation, 2015,29(2):101-125.

[2] GONZÁLEZ-RUBIO J, NAVARRO-CERDÁN, J R, CASACUBERTA F. Dimensionality reduction methods for machine translation quality estimation[J]. Machine translation, 2013,27(3/4):281-301.

[3] SPECIA L, PAETZOLD G, SCARTON C. Multi-level translation quality prediction with QuEst++[C]//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing. Beijing, China:ACL,2015:115-120.

[4] SPECIA L, RAJ D, TURCHI M. Machine translation evaluation versus quality estimation[J]. Machine translation ,2010,24(1): 39-50.

[5] SPECIA L, CANCEDDA N, TURCHI M, et al. Estimating the sentence-level quality of machine translation systems[C]// 13th Annual Conference of the European Association for Machine Translation. Barcelona:EMAT, 2009:28-35.

[6] SPECIA L, SHAH K, De SOUZA J G C, et al. QuEst - A translation quality estimation framework[C]//ACL (Conference System Demonstrations). Sofia, Bulgaria:dblp, 2013:79-84.

[7] SHAH K, LOGACHEVA V, PAETZOLD G H, et al. Shef-nn: Translation quality estimation with neural networks [C]// Proceedings of the Tenth Workshop on Statistical Machine Translation. Lisbon, Portugal: Association for Computational Linguistics, 2015:342-347.

[8] SHAH K, BOUGARES F, BARRAULT L, et al. SHEF-LIUM-NN: Sentence - level quality estimation with neural network features[C]// Proceedings of the First Conference on Machine Translation. Berlin, Germany:ACL,2016:838-842.

[9] CHEN Zhiming, TAN Yiming, ZHANG Chenlin, et al. Improving machine translation quality estimation with neural network features [C]//Proceedings of the Second Conference on Machine Translation. Copenhagen, Denmark: ACL, 2017:551-555.

[10] SHAH K, RAYMOND W M N, BOUGARES F, et al. Investigating continuous space language models for machine translation quality estimation[C]//2015 Conference on Empirical Methods in Natural Language Processing. Lisbon, Portugal:dblp, 2015:1073-1078.

[11] MARTINS A F T, JUNCZYS-DOWMUNT M, KEPLER F N, et al. Pushing the limits of translation quality estimation [J]. Transactions of the Association for Computational Linguistics, 2017,5: 205-218.

[12] KIM H, LEE J H. A recurrent neural networks approach for estimating the quality of machine translation output [C]// Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. San Diego, CA, USA:ACL, 2016: 494-498.

[13] KIM H, JUNG H Y, KWON H S, et al. Predictor-estimator: neural quality estimation based on target word prediction for machine translation[J]. ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP) ,2017,17 (1): 3:1-3:22.

[14] 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. Copenhagen, Denmark:ACL, 2017:562-568.

[15] European Commission. Shared task: Quality estimation[EB/OL]. [2017 - 09 - 07]. http://www. statmt. org/wmt17/quality - estimation-task.html.