小 型 微 型 计 算 机 系 统
Journal of Chinese Computer Systems
DOI:10.20009/j.cnki.21-1106/TP.2022-0333
2023 年 12 月 第 12 期
Vol.44 No.12 2023

融合无监督质量估计指标的译文质量估计方法

胡 雨,朱俊国,余正涛

(昆明理工大学 信息工程与自动化学院,昆明 650500)
(昆明理工大学 云南省人工智能重点实验室,昆明 650500)
E-mail:jg.zhu.hit@qq.com

摘 要: 质量估计的目的是在没有参考译文的情况下衡量翻译内容的质量,这对于需要高质量翻译任务中的机器翻译系统至关重要.针对有监督的翻译质量估计中普遍存在的缺乏标记训练数据和模型框架中两阶段学习目标差异的问题,以及无监督的质量估计中存在的因为任务目标模糊、特征学习不完全而导致最终结果远不如有监督模型的问题,本文提出了将无监督质量估计中的估计指标当作特征融入有监督的质量估计模型中的方法,以此来互相弥补两种模型之间存在的缺点.在 WMT2020 的高资源对和低资源对上的实验结果证实,相对于基线的有监督和无监督模型,两者结合的方法能够更好的提高翻译质量估计的准确性,与人工评分的皮尔逊相关系数都有所提升.

关 键 词: 翻译质量估计;有监督;无监督;特征融合
中图分类号: TP391 **文献标识码:** A **文 章 编 号:** 1000-1220(2023)12-2715-06

Quality Estimation Method Incorporating Unsupervised Quality Estimation Indicators

HU Yu,ZHU Jun-guo,YU Zheng-tao
(Faculty of Information Engineering and Automation,Kunming University of Science and Technology,Kunming 650500,China)
(Yunnan Key Laboratory of Artificial Intelligence,Kunming University of Science and Technology,Kunming 650500,China)

Abstract: The purpose of quality estimation is to measure the quality of translated content without reference translation,which is very important for machine translation systems that need high-quality translation tasks. In view of the widespread lack of labeled training data and the difference between the two-stage learning objectives in the model framework in supervised translation quality estimation,as well as the problem that the final result of unsupervised translation quality estimation is far inferior to that of supervised model because of vague task objectives and incomplete feature learning,this paper puts forward a method of integrating the estimation indicators in unsupervised translation quality estimation into supervised quality estimation model as features,so as to make up for the shortcomings between the two models. The experimental results on WMT2020′s high-resource pair and low-resource pair show that,compared with the baseline supervised and unsupervised models,the combination of the two methods can better improve the accuracy of translation quality estimation,and the Pearson correlation coefficient with manual evaluation has been improved.

Key words: translation quality estimation;supervised;unsupervised;features fusion

1 引 言

随着神经机器翻译(NMT)技术的发展,机器翻译系统已经被应用在很多现实场景中,但是在一些特定场景中机器翻译系统仍然无法避免地产生错误的翻译,尤其是对于没有足够训练数据的低资源语言对[1].如果因为系统翻译错误而导致最终用户有所损失,可能会损害用户对翻译系统的信任,因此,有一个反馈机制来告知用户给定机器翻译输出的可信度是非常重要的,用户可以在知道机器翻译结果质量好坏的情况下,选择决定是否采用它们[2].

质量估计(Quality Estimation,QE)能够为用户提供这样的反馈机制,它旨在没有人工参考译文的情况下,能够对机器翻译系统的输出译文质量进行实时的预测估计,目前,有不同类型的质量估计任务,包括词级、短语级、句子级和文档级[3].

QuEst[4]和 QuEst++[5]等早期的译文质量估计系统严重依赖于语言处理和特征工程,通过从统计机器翻译系统中提取或从源语言和翻译句子中获得大量人工设计的特征,如语言模型概率、长度特征等特征,输入到传统的机器学习算法中估计翻译的质量,这些方法提供了良好的结果,但是适应性较差.随着统计机器翻译(SMT)向神经机器翻译的发展,逐渐被基于深度神经网络的质量估计方法所取代,后者通常很少或者不再依赖于语言处理.例如,在 2017 年 WMT 的质量估计评测任务中表现最好的系统——POSTECH[6],它就是一个纯神经系统,并且完全不依赖于特征工程.

目前,质量估计中表现最好的通用方法是基于知识迁移

收稿日期:2022-05-26 收修改稿日期:2022-07-01 基金项目:国家自然科学基金地区基金项目(62166022)资助;国家自然科学基金重点项目(61732005)资助;云南省科技厅面上项目(202101AT07007)资助;云南省人培项目(KKSY201903018)资助. 作者简介:**胡 雨**,女,1999 年生,硕士研究生,研究方向为自然语言处理;**朱俊国**(通讯作者),男,1982 年生,博士,讲师,CCF 会员,研究方向为自然语言处理;**余正涛**,男,1970 年生,博士,教授,博士生导师,CCF 会员,研究方向为自然语言处理.

的质量估计框架,这一类方法的模型系统通常分为两个阶段:第 1 阶段,预测器(Predictor)是在大量的平行预料上进行预训练的“词预测”模型,用来学习源语句和目标语句中有用的向量表征. 第 2 阶段,估计器(Estimator)在少量的 QE 标记数据上进行训练,学习如何使用在预测器中提取的句对特征上拟合 QE 标记[7]. POSTECH 系统的预测器模型是基于一个双向和双语循环神经网络(BiRNN),从影响每个目标单词预测的神经网络组件中生成质量估计特征向量(quality estimation feature vectors,QEFVs),再将 QEFVs 输入估计器中来产生质量估计. 在此基础上,阿里在 WMT18 的 QE 评测任务中提出了“双语专家”模型[8],特征抽取模型与估计模型使用了 Transformer[9] 与双向 LSTM 的框架,相比于双向 RNN 的词预测,完全基于注意力的 Transformer 框架能够更好地对原文和译文中的信息或特征进行抽取,除了利用抽取到的 QEFVs 外,还利用“双语专家”模型词预测为当前词和强制输出为目标端词之间的差异,人为的设计了四维的错误匹配特征. 尽管取得了很好的结果,但这些方法是基于复杂的神经网络,需要密集的资源训练,并且依赖大量的平行数据和计算资源. 随着以 ELMo[10] 和 BERT[11] 为代表的预训练的语言模型的提出,2019 年 Lample[12] 和 Commeau[13] 等人尝试将预训练的语言模型 BERT 和 XLM 应用到翻译质量估计中,取代了用大量平行语料进行词预测预训练的过程,实验结果表明,这种方法存在一定的潜力. 2020 年 Ranasinghe[14] 等人将 XLM-Roberta (XLM-R)[15] 应用到质量估计任务中,并取得了良好的结果,通过使用微调的跨语言嵌入,避免了使用大量并行语料库进行训练的需要,并减少了复杂神经网络结构的负担.

虽然预测器阶段所需要的大量平行数据和训练的问题得到了缓解,但是估计器阶段需要使用到的人工标记的 QE 数据仍然是一大难题. QE 数据除了真实的原语句和机器翻译的目标句之外,还需要由人工估计或者人工后期编辑产生高质量的注释. 然而在实践中获得这样的注释数据是非常昂贵且耗时的,WMT 质量估计的共享任务中每年所提供的 QE 标记数据集通常只有两三万条左右,相比于成百上千万的平行语料而言,是非常稀少的. 为了解决这个问题,越来越多的学者开始研究不需要标记数据的无监督质量估计方法. Fomicheva[16] 等人提出了一种除了 NMT 翻译系统之外,不需要训练或者获得任何额外的资源的无监督质量估计方法,利用来自 NMT 模型的输出概率分布和注意力权重的熵,还通过不确定性量化的方法让输出概率分布能够更有效的代表翻译输出质量. Tuan[17] 等人使用合成的标注 QE 数据来训练无监督的 QE 模型,通过用机器翻译模型或者掩码语言模型构建合成目标句,并使用 TER 工具将合成目标句和参考译文进行比较,生成高质量的伪标签注释,但是合成数据会产生偏置噪声,包含很多更严重且稀少的错误. 为了缓解合成数据中噪声的影响,Zheng[18] 等人提出了一种自监督的质量估计方法,通过恢复被屏蔽的目标词来进行质量估计,进一步提高了机器预测与人工打分之间的相关性. 虽然无监督的质量估计方法解决了标注数据稀少的问题,但是整体的结果相比于有标记数据学习的方法,还是存在的很大差距.

前面所介绍的基于知识迁移的两阶段 QE 框架 Predictor-Estimator,证实了通过预测器在平行语料上的训练学习,可以有助于估计器阶段的效果,但两阶段由于训练目标的差异:预测器在平行语料上做词预测的任务,估计器在标记数据上预测翻译的质量. 这会导致大规模的平行语料没有被充分学习利用,因为预测器的训练过程与最终的质量估计任务目标存在差异,从而学习不到 QE 任务实际需要的知识.

考虑到无监督质量估计方法主要是使用机器翻译系统的内部信息作为指标,也就是直接针对最终的目标 QE 任务对源语言和目标语句进行学习预测,因此有可能将其作为特征直接加入到估计器的阶段,弥补预测器阶段因为目标差异学习不充分的问题. 对此,本文尝试将两种质量估计的方法结合到一起,以达到互相弥补问题,提升性能的效果. 在这方面,本文尝试使用上述想法作为实验的起点,实验结果表明,引入特定的无监督 QE 指标的方法优于不引入该指标的方法.

2 相关工作

2.1 无监督 QE 的估计指标

正如上文中提到的,Fomicheva[16] 等人设计了一种直接从 NMT 系统中提取有用信息作为质量估计指标,在无监督 QE 方法中取得了最好的效果,甚至可以和有监督的 QE 模型相媲美. 作为翻译的副产品,除了机器翻译系统本身之外,该方法不需要训练或者额外的资源,因此本文中将其中的估计指标提取出来作为特征用于后面的任务.

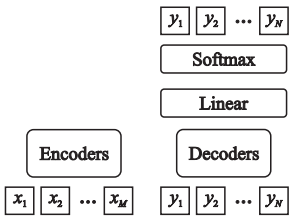

图 1 神经机器翻译模型

Fig. 1 Neural machine translation model

图 1 展示了一个使用注意力机制的序列到序列的神经机器翻译模型,同时也是无监督质量估计的估计指标抽取的模型. 给定输入序列 $x = x_1, x_2, \cdots, x_M$,假设解码器生成输出序列 $y = y_1, y_2, \cdots, y_N$,输出序列长度为 N,生成 y 的概率可以被分解为:

$$p(y|x,\theta) = \prod_{n=1}^{N} p(y_n|y_{<n},x,\theta) \tag{1}$$

其中 θ 代表模型参数,系统词汇的概率分布 $p(y_n|y_{<n},\theta)$ 由解码器在每个时间步长通过 Softmax 函数产生.

本文采用 Fomicheva 等人[16] 的论文中的 3 个无监督质量估计指标来进行特征融入:

· 首先是对词级的概率分布取对数,再按照序列长度取平均值得到序列级翻译概率:

$$\text{Seq} = \frac{1}{N}\sum_{n=1}^{N} \log p(y_n|y_{<n},x,\theta) \tag{2}$$

· 使用 Monte Carlo dropout(MC dropout)[19],通过将神经元随机屏蔽为零,用于减少 NMT 模型的过拟合,假设通过此不确定性量化的方法,模型的参数 θ 受到干扰变为 $\hat{\theta}$. 对每个序列使用 MC dropout 进行 T 次推理,并计算 T 次的序列概率的期望值:

$$\text{MC-Seq} = \frac{1}{T}\sum_{t=1}^{T}(\text{Seq})_{\hat{\theta}} \tag{3}$$

·当通过 MC dropout 进行不确定性量化时,对同一个源语句会产生不同的翻译输出,假设机器翻译不同输出之间的差异也可能代表了不确定性以及源语句的复杂性,计算了一组翻译输出 H 之间的平均相似性分数:

$$\text{MC-Sim} = \frac{1}{C}\sum_{i=1}^{|H|}\sum_{j=1}^{|H|} sim(h_i, h_j) \tag{4}$$

2.2　基线模型

TransQuest 是 WMT2020 的质量估计任务中表现最好的开源 QE 框架,通过将多语言预训练模型 XLM-R 应用到质量估计任务中,提出了一个简单的 QE 框架,可以很容易地训练不同语言对或者不同领域语言的输入,实现不同资源语言对之间的迁移学习[14]. 其中 XLM-R 模型是在一个大规模的多语言数据集上进行训练:从 CommonCrawl 数据集中提取 104 种语言的过滤文本,在共 2.5TB 的数据上仅使用 RoBERTa 的掩码语言模型(masked language model,MLM)目标进行训练. 该框架在跨语言问答、分类推理等任务上都取得了出色的结果,同时在 QE 任务上也获得了更好的效果.

其中的一种框架如图 2 的左半部分所示,将原文及其译文的连接作为输入序列,然后通过 XLM-R 模型获取质量估计的输出表征. 这样的方式使得模型的训练计算强度要小很多,因为框架不依赖于复杂的输出层,但是因为依赖于预训练的 XLM-R 模型,模型也存在一些不足. 首先,用于训练 XLM-R 的是多语言的单语语料,会导致模型在双语语义相关方面的学习能力有所欠缺,而且在下游的 QE 任务中,用于训练的 QE 标记数据集数量少,会让最终输出的 QE 表征内容学习不完整. 对此,可以尝试将前面的无监督质量估计指标作为特征直接加入到有监督 QE 的估计器的阶段,弥补预测器阶段 XLM-R 模型学习不充分的问题.

3　融合无监督质量估计指标的质量估计方法

受 TransQuest 工作的启发,本文所提出的结合有监督 QE 和无监督 QE指标的模型框架如图2所示. 首先采用2.1节

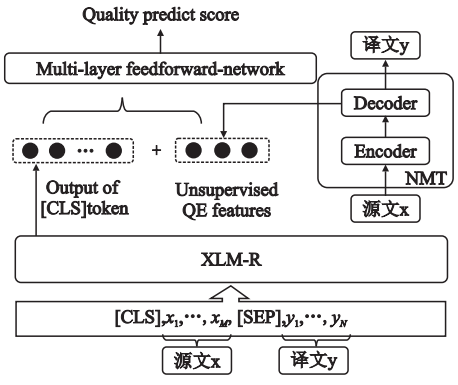

图 2　模型整体框架

Fig. 2　Model overall framwork

中所述的无监督质量估计方法,在 NMT 模型的翻译过程中直接从源文 x 和译文 y 中提取出评价指标 UnQE:Seq、MC-Seq、MC-Sim,上述 3 种评价指标的维度数都为 1,即每个句子都对应一个分数;再将这 3 种评价指标当作特征分别融入或组合融入到有监督 QE 中进行特征融合.

其次如图 2 左边所示,模型输入为源文 x 及其翻译 y 的串联拼接:$S = \{[\text{CLS}], x_1, \cdots, x_M, [\text{SEP}], y_1, \cdots, y_N\}$ 序列的第 1 个标记为[CLS],源文 x 和译文 y 中间由标记分隔,通过基于 Transformer 的预训练模型 XLM-R,将序列 S 转换输出为融合上下文信息的词嵌入向量. 序列的第 1 个标记始终为[CLS],它包含表示整个序列的特殊词嵌入,是代表正在输入的整个序列的一个很好的标记,之后是序列中每个 token 的词嵌入. 将序列 S 通过 XLM-R 模型转换后的整个输出切片获得单独的[CLS]token 的表征:

$$[\text{CLS}]_{embdding} = \text{XLM-R}(S) * [:, 0, :] \tag{5}$$

选取[CLS]标记的词嵌入向量作为整个序列的质量估计特征向量,然后在此时将无监督估计指标 UnQE 融合进去,再将整体的特征向量输入回归器预测序列的质量分数,即输入到具有隐藏层和 *tanh* 激活函数的多层前馈神经网络中,使用均方误差损失作为目标函数,然后多维度的张量就会被投射到维度为 1 的张量中,进而作为译文质量估计分数:

$$\text{QEScore} = tanh(W * [\text{CLS}]_{embdding} \oplus \text{UnQE}] + b) \tag{6}$$

其中,⊕表示两种张量进行连接的操作,W、b 参数都是模型训练中学习到的,分别表示权重和偏置的张量,为了能够方便估计翻译的质量,激活函数将特征向量映射到[-1,1]的范围内.

4　实验与分析

4.1　实验设置

QE 标记数据:直接估计(Direct Assessment,DA)[20]:与基于机器翻译和参考译文之间的编辑距离的人工翻译错误率(Human-Targeted Translation Error Rate,HTER)[21]不同,每个句子的 DA 分数由专业的翻译人员进行注释. 根据至少 3 名专业翻译人员对机器翻译质量进行直接的感知判断,每一句机器译文都被在 0 ~ 100 的范围内进行一个打分,然后将 DA 分数转换为 z-scores 标准化. 尽管 HTER 是通用的翻译质量估计指标,但研究人员对这一指标的可靠性提出了质疑[22],而 DA 方法已经被证明可以提高人工估计的再现性,在估计质量估计系统的性能时更加可靠,为自动估计指标提供了更可靠的标准. 本文使用发布于 WMT2020 质量估计任务 1 中的数据集,表 1 为中文-英文数据集里面的一个 DA 数据例子.

表 1　人工直接打分的数据例子

Table 1　Example of direct assessment data

Original	Many butterflies flutter among the flowers and grasses.
Translation	许多蝴蝶在花草间飘动.
Scores	[50,75,58,93,90,95]
Mean	76.833333
z-scores	[-0.5485432, 0.5715456, 0.7163991, 1.1256175, 1.0690328, 1.4257850]
z-mean	0.7266395

模型设置:对于无监督质量估计特征指标提取部分,NMT 模型的翻译质量受到训练数据集数量的强烈影响,同时也会影响提取的指标. 为了区分差异,本文对两类语言对进行了实验:高资源语言对,包括英语-德语(English-German,En-De)

和英语-汉语(English-Chinese,En-Zh),低资源语言对,包括僧伽罗语-英语(Sinhala-English,Si-En)和尼泊尔语-英语(Nepali-English,Ne-En).其中的高资源和低资源的区别为用来训练 NMT 模型的平行语料数据集的多少,使用开源的 fairseq 工具为 4 组语言对训练了基于 Transformer 的 NMT 模型.针对高资源语言对的 NMT 模型,采用标准的 Transformer 架构;而低资源语言对的 MT 系统采用基于 Vaswani 等人[9]在论文中定义的 Big Transformer 架构进行训练,模型按照 FLORES 半监督设置[23]进行训练:训练一个反向 MT 系统,使用该系统将单语目标句子翻译为源语言;然后将有噪音(反向翻译)的源句子和原始目标句子合并,并将它们作为额外的并行数据添加到正向的 MT 系统中.其中包括使用源语言和目标句子进行两次迭代回译,表 2 为实验中所使用数据的具体数量信息,其中回译所使用的单语数据即为表中低资源语言对的目标端数据.

表 2 实验数据集信息

Table 2 Information of experimental data sets

	平行语料规模	QE 句对数量 源语言端词数量 训练集/验证集/测试集
En-Zh	2.75G - 2.5G	7000/1000/1000 115,585/16,307/16,765
En-De	2.58G - 2.83G	7000/1000/1000 114,980/16,519/16,371
Si-En	49.6M - 21.1M	7000/1000/1000 109,515/15,708/15,821
Ne-En	38.4M - 13.3M	7000/1000/1000 104,934/15,144/14,770

在无监督模型中,对于基于 MC dropout 的指标,本文使 T = 10 个推理过程来获得后验概率分布,对于每一组语言对中的 DA 标记数据,都包含 7k 条的 QE 标记数据用于训练,1K 条用于验证以及 1K 条用于测试,数据都来源于 WMT2020 中 QE 任务的公开数据集.此外,为了能够结合无监督质量估计特征,有监督的质量估计部分使用相同的 QE 标记数据集进行训练,batchsize 设为 8,使用 Adam 优化器,学习率设置为 2e-5,同时也设置 dropout 防止过拟合(rate = 0.1).上述关于无监督质量估计的模型和数据在 Fomicheva 等人的论文中提供[1].

4.2 实验结果

特征融合:为了融合有监督的特征向量$[\mathrm{CLS}]_{embdding}$和无监督的估计指标UnQE从而得到最终的质量估计向量$\mathbf{H}$,本文尝试在融合单独的无监督质量指标时对比了以下两种融合策略:

· Concatenation - 将向量直接按维度进行串联拼接

$$\mathbf{H} = [[\mathrm{CLS}]_{embdding}; \mathrm{UnQE}] \tag{7}$$

· Multiplication - 将向量互乘:

$$\mathbf{H} = [[\mathrm{CLS}]_{embdding} * \mathrm{UnQE}] \tag{8}$$

为了比较上述提出的两种不同融合策略对最终性能的影响,所以在特征融合的 QE 框架中,本文在中英数据集上进行了一组对比实验,结果如表 3 所示,结果为基线模型采用两种不同的融合策略和无监督 QE 指标 MC-seq 进行融合.

表 3 两种融合策略的结果

Table 3 Result of two fusion strategies

Strategy	Pear	Spear	MAE	RMSE
Mult	0.5294	0.5134	0.8076	0.9766
Conc	0.5458	0.5512	0.7925	0.9618

为了估计翻译质量估计系统的性能,通常使用皮尔逊相关系数(Pearson correlation coefficient)、斯皮尔曼相关系数(Spearman correlation coefficient)、平均绝对误差(MAE)和均方根误差(RMSE).皮尔逊相关系数是体现了人为打分与机器自动打分之间线性相关的重要评价指标.同时为了体现这两个打分向量的单调性走向分布是否一致,使用了斯皮尔曼相关系数来估计自动打分排名与人工打分排名之间的排名分布相关性.MAE 和 RMSE 作为协助评价指标,与以上所述这两个评价指标不一样,它们的分数越低,代表系统性能越好.

从表 3 中的结果可以看出,串联拼接的融合策略优于直接相乘的融合策略,这可能是因为当 1024 维的隐藏层向量乘以单维的特征向量时,得到的结果向量中携带的隐式信息重叠,导致融合的结果向量的特征冗余.然而,串联拼接的融合方法根据维数将每个向量直接拼接,该方法不修改每个特征的任何维度内容,使每个特征所携带的信息得到充分表达,特征信息量之间的增加互补使得特征融合的效果更加理想.因此,在下文中,仅对使用效果良好的串联拼接的融合方法的系统性能进行了比较和分析,包括同时融合多个无监督评价指标时,都采用串联拼接的方法.

最终结果:为了比较无监督质量估计框架中抽取出来的单一特征和不同特征组合对翻译质量估计的影响,本文在控制变量条件下进行了几组对比实验,并根据实验结果分析了每个特征对整体性能的影响.在表 4 中,baseline 表示仅使用有监督的TransQuest基线模型,+ 表示基于基线模型所融合

表 4 DA 和机器自动估计分数之间的皮尔逊相关和斯皮尔曼相关结果

Table 4 Result of pearson correlation and spearman correlation between DA and machine automatic evaluation score

methods	Pearson Correlation				Spearman Correlation			
	Low-resource		High-resourse		Low-resource		High-resourse	
	Si-En	Ne-En	Si-En	Ne-En	Si-En	Ne-En	Si-En	Ne-En
Baseline	0.6362	0.7866	0.4353	0.4875	0.6092	0.7656	0.4601	0.4813
+ Seq	0.6133	0.7805	0.4775	0.5173	0.5905	0.7544	0.5138	0.5176
+ MC-Seq	**0.6520**	**0.8022**	**0.4841**	**0.5458**	0.6178	**0.7783**	0.5171	**0.5512**
+ MC-Sim	0.6364	0.7823	0.4686	0.4948	0.6139	0.7566	0.4998	0.4906
+ MC-Seq + MC-Sim	0.6512	0.7857	0.4765	**0.5477**	**0.6286**	0.7657	**0.5236**	0.5441

[1] https://github.com/facebookresearch/mlqe/

的特征名称. 表中给出了人工打分 DA 和机器自动估计分数之间的皮尔逊相关性和斯皮尔曼相关性的结果. 在对 4 种语言对的实验中,4 种不同的无监督 QE 特征的不同组合都进行了融合实验. 与没有融合特征的基线相比,整合的结果在所有 4 组语言对中都取得了一定的改进,不管是单独还是组合融入. 尤其是在单独融合 MC-Seq 时,结果的提升幅度是最大.

4.3　实验分析

不同的组合融入:忽略不同语言对之间的差异,仅针对不同的特征组合融合效果来看,也就是通过纵向观察表 4 中的最终效果. 当单独融合集成 MC-Seq 时,基线的提升效果最大,最终机器对译文质量估计的预测精准度最高. 通过具体分析,可以发现,在皮尔逊相关系数方面,添加 MC-Seq 后的系统优于仅集成了 Seq 或者 MC-Sim 的系统.

其中 MC-Seq 优于 Seq 的原因是因为在获取 MC-Seq 的过程中使用了不确定性量化的 MC-dropout,这使得在神经网络的训练过程中,对迭代中的某一层神经网络,随机选择并暂时隐藏屏蔽其中的一些神经元,然后进行训练和优化,在下一次迭代中,继续随机隐藏屏蔽一些神经元,直到训练结束. 由于神经元被随机隐藏屏蔽为零,每个批次都在训练不同的神经网络. 通过这种方式,模型的泛化性变的更强,减少隐藏层节点之间的交互作用,因为它不会太依赖于某些局部特征,并且可以有效的减少过拟合现象. 使用 dropout 后,模型的不确定性量化提高了机器打分与人工打分的相关性,与融合 Seq 相比,融合 MC-Seq 所携带的信息能更准确地反映译文与源语言之间的映射关系,从而提升融合后模型对译文质量估计的准确性.

MC-Sim 指标是衡量通过干扰模型参数产生的最大似然假设之间的可变性,也就是同一源文产生的不同译文之间的相似性. 尽管这一指标可以估计无监督质量估计中的译文质量,但这一特征并没有显著改善提升基线的效果,其中包含的特征信息不能在基线模型的隐藏特征向量基础上提高译文质量估计的效果. 由于 Seq 和 MC-Seq 之间的信息冗余,两者都是译文序列的概率分布期望值,但 MC-Seq 的效果更好,所以选择了两者中性能更好的 MC-Seq 并与 MC-Sim 相互组合拼接融入. 与单独仅集成融合 MC-Sim 相比,结果有所改善,但是略差于单独集成融合 MC-Seq,这也验证了单独集成融合 MC-Seq 的改善效果最好,且组合比单独集成融合的效果更差.

图 3 中的两个散点图显示了验证集中 1000 个句子的机器自动估计的分数与人工评分 DA 评分之间的相关性. 以语言对 En-Zh 为例,图 3 上边显示的是基线模型,没有融入任何特征,图 3 下边显示的是融入 MC-Seq 后的分布. 两个散点图中的人工 DA 评分(z_mean)都是一样的,因此其在散点图中的分布是固定的.

通过比较机器自动估计打分(predicted z_mean)在两个图中的分布,可以发现,通过融入添加特征 MC-Seq 后,predicted z_mean 的分布更接近 z_mean,更加收敛于 z_mean. 这进一步证明,添加 MC-Seq 后模型获得的预测分数与人工的 DA 分数具有更高的相关性,判断更准确.

不同资源的语言对:通过观察 4 种语言对之间的结果,高资源语言对和低资源语言对的最终结果也有很大差异. 与基线相比,融合 MC-Seq 方法的 Pearson 相关系数在 En-De 和 En-Zh 的高资源语言任务中分别增加了 4.88% 和 5.83%,而在 Si-En 和 Ne-En 的低资源语言任务中仅增加了 1.58% 和1.56%. 虽然相同的特征对不同的语言对都有着促进的作用,但影响的程度是不同的. 本文猜测,是因为高资源语言对的翻译句子质量优于低资源语言对(有着较高的人工 DA 平均分数).

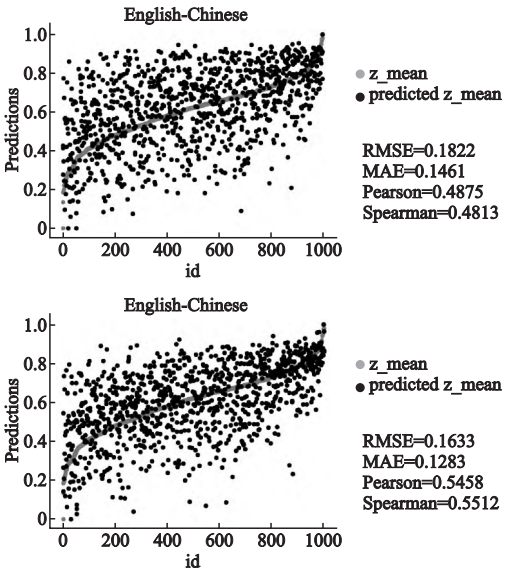

图 3　基线(上)和 + MC-Seq(下)的机器自动估计打分(predicted z_mean)和人工评分(z_mean)之间的相关性散点图

Fig. 3　Scatter plots for the correlation(En-Zh) between machine automatic evaluation score(predicted z_mean) and standardized DA score(z_mean) for baseline(up) and + MC-Seq(down)

表 5 中的数据就是为了验证这个猜想. 在融合 MC-Seq 特征后,本文对 4 种语言对的最终结果进行了再分化实验. 本文将验证集中的 1000 个句子分为两部分:z-mean 大于 0 和 z-mean 小于 0,其中 z-mean 的数值越大,代表句子的人工平均 DA 分数越高,句子的整体质量越好. 表中的两行数据分别代表句子数和皮尔逊相关系数. 从表中的数据可以看出,当 z-mean 大于 0 时,4 种语言对的 Pearson 系数都高于 z-mean 小于 0 的 Pearson 系数,这表明特征 MC-Seq 更有助于预测句子质量更好的最终翻译.

表 5　以 z-mean 为界限,将 1000 条句子分为两部分

Table 5　Taking z-mean = 0 as the boundary, the 1000 sentences set are divided into two parts

z-mean	Low-resourse		High-resourse	
	Si-En	Ne-En	Si-En	Ne-En
>0	493	416	564	617
	0.5148	0.6631	0.3186	0.4348
0 <	507	584	436	383
	0.5142	0.5310	0.2247	0.3088

在高资源语言对 En-De 和 En-Zh 中,z-mean 大于 0 的句子数分别为 567 和 617,远远超过在低资源语言对 Si-En 和 Ne-En 中的 493 和 416,这也证明了本文前面的猜想,因为在高资源语言中,译文质量好的句子比在低资源语言中要多,因

此,更有助于预测高资源语言对的最终译文的质量.

5 结 论

在句子层面的翻译质量估计任务中,本文从无监督的质量估计系统开始,在无监督方法中,从 NMT 系统中提取的指标可能包含丰富的信息源,这些信息源使用了一系列基于 NMT 模型的 softmax 输出概率分布而变形的指标,它可以弥补有监督 QE 模型中预测器阶段对语料学习不充分的问题,补充加入源文本与对应翻译之间映射信息. 针对这种情况,本文提出了一种质量估计方法,该方法以无监督译文质量评价的估计指标为特征,并将其融合集成到有监督译文质量评价中. 实验结果表明,该方法提高了句子级翻译质量估计任务中机器自动评分与人工评分之间的相关性. 同时,这项工作可以以各种方式扩展,首先,这一思想可以应用于不同级别的质量估计模型,例如单词级别或文档级别;其次,其他无监督的质量估计指标也可以进行组合和融入.

References:

[1] Barrault L, Biesialska M, Bojar O, et al. Findings of the 2020 conference on machine translation[C]//Proceedings of the 5th Conference on Machine Translation, 2020: 1-55.

[2] Castilho S, Moorkens J, Gaspari F, et al. Is neural machine translation the new state of the art? [J]. The Prague Bulletin of Mathematical Linguistics (PBML), 2017, 108: 109-120.

[3] Ive J, Blain F, Specia L. Deepquest: a framework for neural-based quality estimation[C]//27th International Conference on Computational Linguistics, 2018: 3146-3157.

[4] Specia L, Shah K, Camargo J G, et al. QuEst-a translation quality estimation framework[C]//Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, System Demonstrations, 2013: 79-84.

[5] Specia L, Paetzold G, Scarton C. Multi-level translation quality prediction with QuEst + +[C]//Proceedings of Association for Computational Linguistics and the Asian Federation of Natural Language Processing, 2015: 115-120.

[6] 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 2nd Conference on Machine Translation, 2017: 562-568.

[7] Kim H, Jung H Y, Kwon H, 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, 2017, 17(1): 1-22.

[8] Kreutzer J, Schamoni S, Riezler S. Quality estimation from ScraTCH (QUETCH): deep learning for word-level translation quality estimation [C]//10th Workshop on Statistical Machine Translation, 2015: 316-322.

[9] Vaswani A, Shazeer N, Parmar N. et al. Attention is all you need [C]//Proceedings of the 31st International Conference on Neural Information Processing Systems, 2017: 6000-6010.

[10] Peters M, Neumann M, Iyyer M, et al. Deep contextualized word representations[C]//Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics, 2018: 2227-2237.

[11] Devlin J, Chang M W, Lee K, et al. BERT: pretraining of deep bidirectional transformers for language understanding[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), 2019: 4171-4186.

[12] Kim H, Lim J H, Kim H K, et al. QE BERT: bilingual BERT using multitask learning for neural quality estimation[C]//Proceedings of the 4th Conference on Machine Translation (WMT), 2019: 85-89.

[13] Kepler F, Trénous J, Treviso M, et al. Unbabel′s participation in the WMT19 translation quality estimation shared task[C]//Proceedings of the 4th Conference on Machine Translation (WMT), 2019: 78-84.

[14] Ranasinghe T, Orasan C, Mitko V R. Trans Quest: translation quality estimation with crosslingual transformers[J]. arXiv preprints arXiv: 2011.01536, 2020.

[15] Conneau A, Khandelwal K, Goyal N, et al. Unsupervised crosslingual representation learning at scale[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL), 2019: 8440-8451.

[16] Fomicheva M, Sun S, Yankovskaya L, et al. Unsupervised quality estimation for neural machine translation[J]. Transactions of the Association for Computational Linguistics, 2020, 8(1): 539-555.

[17] Tuan Y L, El-Kishky A, Renduchintala A, et al. Quality estimation without human-labeled data[C]//Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, 2021: 619-625.

[18] Zheng Y, Tan Z, Zhang M, et al. Self-supervised quality estimation for machine translation[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2021: 3322-3334.

[19] Gal Y, Ghahramani Z. Dropout as a Bayesian approximation: representing model uncertainty in deep learning[J]. arXiv preprints arXiv: 1506.02142, 2015.

[20] Graham Y, Baldwin T, Moffat A, et al. Can machine translation systems be evaluated by the crowd alone[J]. Natural Language Engineering, 2017, 23(1): 3-30.

[21] Snover M, Dorr B, Schwartz R, et al. A study of translation edit rate with targeted human annotation[C]//Proceedings of the 7th Conference of the Association for Machine Translation in the Americas, Technical Papers, 2006: 223-231.

[22] Graham Y, Baldwin T, Mathur N. Accurate evaluation of segment-level machine translation metrics[C]//Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics, Human Language Technologies, 2015: 1183-1191.

[23] Guzmán F, Chen P J, Ott M, et al. The FLoRes evaluation datasets for low-resource machine translation: Nepali-English and Sinhala-English[C]// Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, 2019: 6098-6111.