计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.2017.20170182
54(8):1804-1812,2017

基于神经网络特征的句子级别译文质量估计

陈志明　李茂西　王明文
(江西师范大学计算机信息工程学院　南昌　330022)
(qqchenzhiming@jxnu.edu.cn)

Sentence-Level Machine Translation Quality Estimation Based on Neural Network Features

Chen Zhiming, Li Maoxi, and Wang Mingwen
(*School of Computer Information Engineering, Jiangxi Normal University, Nanchang* 330022)

Abstract　Machine translation quality estimation is an important task in natural language processing. Unlike the traditional automatic evaluation of machine translation, the quality estimation evaluates the quality of machine translation without human reference. Nowadays, the feature extraction approaches of sentence-level quality estimation depend heavily on linguistic analysis, which leads to the lack of generalization ability and restricts the system performance of the subsequent support vector regression algorithm. In order to solve this problem, we extract sentence embedding features using context-based word prediction model and matrix decomposition model in deep learning, and enrich the features with recurrent neural network language model feature to further improve the correlation between the automatic quality estimation approach and human judgments. The experimental results on the datasets of WMT'15 and WMT'16 machine translation quality estimation subtasks show that the system performance of extracting the sentence embedding features by the context-based word prediction model is better than the traditional QuEst method and the approach that extracts sentence embedding features by the continuous space language model, which reveals that the proposed feature extraction approach can significantly improve the system performance of machine translation quality estimation without linguistic analysis.

Key words　machine translation quality estimation; sentence-level; word embedding; recurrent neural network language model; support vector regression

摘　要　机器翻译质量估计是自然语言处理中的一个重要任务,与传统的机器翻译自动评价方法不同,译文质量估计方法评估机器译文的质量不使用人工参考译文.针对目前句子级别机器译文质量估计特征提取严重依赖语言学分析导致泛化能力不足,并且制约着后续支持向量回归算法的性能,提出了利用深度学习中上下文单词预测模型和矩阵分解模型提取句子向量特征,并将其与递归神经网络语言模型特征相结合来提高译文质量自动估计与人工评价的相关性.在 WMT'15 和 WMT'16 译文质量估计子任务数据集上的实验结果表明:利用上下文单词预测模型提取句子向量特征的方法性能统计一致地优于传统的 QuEst 方法和连续空间语言模型句子向量特征提取方法,这揭示了提出的特征提取方法不仅不需要语言学分析,而且显著地提高了译文质量估计的效果.

收稿日期:2017-03-20;**修回日期**:2017-05-16
基金项目:国家自然科学基金项目(61462044,61662031,61462045)
This work was supported by the National Natural Science Foundation of China (61462044, 61662031, 61462045).
通信作者:李茂西(mosesli@jxnu.edu.cn)

关键词 机器翻译质量估计;句子级别;词向量;递归神经网络语言模型;支持向量回归

中图法分类号 TP391

机器译文质量估计(quality estimation, QE)利用机器学习算法,在没有人工参考译文的情况下自动评价机器翻译的质量.它是统计机器翻译近几年来新兴起的一个研究方向.机器译文质量估计不仅为最终用户提供一个度量译文可用程度的指标,而且可以辅助专业人工译员进行译文的后编辑.因此,它在促进机器翻译技术快速发展和推广应用中起着重要的作用.

在没有人工参考译文对照的情况下,如何定量评价机器译文的质量呢?受语音识别中计算词的置信得分(confidence estimation)的启发,初期机器译文质量估计主要集中于估计译文中词语的置信度[1-2].与估计词语级别译文质量相比,估计句子或系统级别的译文质量更具有实际意义.Blatz 等人把它看作是一个机器学习的 2 类分类问题,通过使用朴素贝叶斯分类器和多层感知机算法,引入 4 类不需要人工参考译文就能提取的 91 个特征来区分机器译文是否正确[3].Quirk 提出利用线性回归算法对机器译文的质量进行分类[4].宁伟等人提出使用浅层词法特征和深层句法特征,利用支持向量机建立模型对译文质量的"好"与"差"进行估计等[5-6].

早期的工作由于对译文质量的分类标准不一致,提取的特征过多,且提取算法与目标语言种类相关,缺乏通用性,因此并没有引起研究者们足够的重视.直到 Specia 等人在前人工作的基础上,提出了译文质量估计方法 QuEst[7],并发布了相关工具包供 WMT QE 句子级别子任务作基线系统.QuEst 方法把机器译文质量估计看作是一个机器学习中的回归问题,从翻译难度、生成的译文流利度和忠实度 3 个方面抽取描述译文质量的特征,利用基于径向基函数核的支持向量回归算法估计机器译文的质量.

围绕机器译文质量估计的 QuEst 方法,研究者们进行了许多卓有成效的工作.这些研究工作主要集中在 2 个方面:1)对机器译文质量估计中机器学习算法的研究.由于在机器译文质量估计中一般提取的特征较多,特征之间存在一定的重叠或者互相依赖.因此,首先要选择相关的特征,Rubino 等人使用回归树学习进行特征的选择;在特征选择之后使用机器学习算法对译文质量进行估计[8].Soricut 等人使用 M5P 模型学习决策树来进行译文质量估计[9];Hardmeier 等人使用基于多项式核的支持向量回归算法来进行译文质量估计[10];Almaghout 和 Specia 使用 Logistic 回归进行译文质量的估计[11].2)对机器译文质量估计中特征的研究.由于缺乏人工参考译文,许多研究工作尝试对机器译文进行深层次语言学分析来提取更多与译文质量密切相关的特征,包括对机器译文进行词性标注[9]、概率上下文无关文法分析[12]、组合范畴文法分析[11]等.

尽管这些方法提高了机器译文质量估计与人工评价的相关性,但是它们采用的还是机器学习中传统的"特征工程+任务建模"的范式.这导致特征提取严重依赖语言学分析模块,特征提取方法与语言种类相关缺乏通用性,并且译文质量估计的效果不甚理想.针对这个问题,本文探索结合深度学习中词语的向量表示和译文的递归神经网络语言模型概率作为特征来进行译文的质量估计.在特征提取中,本文利用大规模单语语料训练词向量和语言模型,因此不需要语言学分析且独立于具体语言.进一步,通过实验验证本文方法的性能优于传统的 QuEst 方法和基于连续空间语言模型的特征提取方法.

1 相关工作

近年来,深度学习在自然语言处理中取得了极大的成功,包括神经网络语言模型的提出[13],神经机器翻译编码解码框架的提出等[14-15].因此,一些工作尝试将其引入到机器译文质量估计任务中以提高译文质量自动估计与人工估计的相关性.

从评价粒度来说,机器翻译质量估计一般分为词级别、句子级别和文档级别,深度学习方法在各级别都有应用.在词语级别机器译文质量估计中,Shah 等人将词向量用做特征以区分机器译文中词语翻译的"好"与"差"[16].Kreutzer 等人将深度前馈神经网络用于词级别的质量估计[17].Patel 等人将递归神经网络语言模型用于词级别质量估计任务[18].在文档级别机器译文质量估计中,Scarton 等人结合篇章分析信息和词向量特征对篇章翻译质量进行估计[19].尽管他们使用词向量作为特征,但是本文方法与其区别在于,本文是在句子级别机器译文质量估计中将句子中词语的向量转化为句子的整体向量,并将其与递归神经网络语言模型结合作为特征.

在句子级别机器译文质量估计中,Shah 等人 2015 年提出利用连续空间语言模型[20]分别训练源语言句子和目标语言句子的语言模型概率用作特征,并融合传统的 QuEst 方法提取的基准特征,来提高译文自动估计与人工评价的相关性[21]. 在 WMT'16 QE 子任务中,Shah 等人在上述工作的基础上,进一步提出增加源语言句子和目标语言句子的交叉熵和句子向量等特征对其进行扩展[22],在提取交叉熵和句子向量特征时,他们利用的仍然是连续空间语言模型. 有部分研究者利用神经网络建立质量估计模型,直接预测机器译文的质量. 例如,Paetzold 等人提出使用多层的 LSTM 网络建立质量估计模型[23]. Kim 等人在基于注意力机制的神经机器翻译编码解码框架[24]的基础上,通过在解码器端增加一层后向 RNN 网络进行机器译文质量估计[25].

本文在 Shah 等人[21-22]的工作基础上进行研究,由于 Shah 等人提取神经网络特征使用的是连续空间语言模型,它是一种前馈神经网络并且输入是固定长度的词序列,不能够很好地处理序列数据;而且该模型使用了多个隐层,随着神经网络的隐层增多和其中节点数量的增加,神经网络的参数将急剧增加导致算法异常复杂. 因此,本文提出分别使用上下文单词预测模型[26]和矩阵分解模型[27]训练词向量进而得到句子向量特征,并将提取的句子向量特征与递归神经网络语言模型概率特征进行结合. 在 WMT'15 QE 和 WMT'16 QE 子任务数据集上[28-29],将上下文单词预测模型和矩阵分解模型提取的句子向量特征与 Shah 等人提出的利用连续空间语言模型提取的句子向量特征进行了对比,实验结果表明:本文提出的方法显著提高了译文质量自动估计的性能.

2 模型和性能评价指标

句子级别机器译文质量估计的目标是给定源语言句子 S 和它的机器译文 T,定量估计机器译文的翻译质量. 假设给定一个训练集 D,它包含 m 个源语言句子和其对应机器译文,以及人工对机器译文的质量评价结果(根据专业译员对机器译文后编辑的次数计算出的 HTER[30]值)$y_i(i=1,2,\cdots,m)$,它可以表示为 $D=\{(S_1,T_1,y_1),(S_2,T_2,y_2),\cdots,(S_m,T_m,y_m)\}$. 通过从源语言句子和其对应的机器译文中抽取描述翻译质量的特征 $\boldsymbol{X}_i(i=1,2,\cdots,m)$,训练集可以进一步表示为 $D'=\{(\boldsymbol{X}_1,y_1),(\boldsymbol{X}_2,y_2),\cdots,(\boldsymbol{X}_m,y_m)\}$. 我们希望在训练集上训练一个函数 f,它在训练集的所有样本上预测损失最小,并且对于在未知样本上提取的特征向量 $\boldsymbol{X}$,$f(\boldsymbol{X})$输出一个反映翻译质量的实值 y. 这实际上是一个回归问题.

本文采用机器学习中经典方法支持向量回归(support vector regression)来进行模型训练和测试. 我们也尝试了复杂的回归模型,如梯度提升回归(gradient boosting regression)和随机森林回归(random forest regression)等,它们增加了模型的复杂度,但是并没有显著提高译文质量估计的性能. 在实验中,支持向量回归核函数选用的是径向基函数,利用格点搜索(grid search)算法和 3 折交叉验证选择模型最优的参数,包括 C,ε,γ.

为了评价机器译文质量估计模型的性能,皮尔森相关系数(Pearson correlation coefficient)r 被用来测定机器译文质量自动估计与人工评价的打分相关性:

$$r=\frac{\sum_{i=1}^{N}(H(T_i)-\overline{H})(V(T_i)-\overline{V})}{\sqrt{\sum_{i=1}^{N}(H(T_i)-\overline{H})^2\sum_{i=1}^{N}(V(T_i)-\overline{V})^2}},$$

其中,$H(T_i)$和 $V(T_i)$分别为机器译文 T_i 的质量自动估计得分和人工打分;$\overline{H}$ 和 $\overline{V}$ 分别是它们的均值;N 表示测试集中机器译文总数目. 皮尔森相关系数越大,自动估计与人工评价结果越吻合. 同时也提供平均绝对误差(mean absolute error, MAE)、均方根误差(root mean squared error, RMSE)作为打分相关性的参考指标.

遵循 WMT'16 QE 子任务的官方评价方法[29],斯皮尔曼相关系数(Spearman correlation coefficient)ρ 被用来测定机器译文质量排名与人工评价排名的排名相关性:

$$\rho=1-\frac{6\times\sum_{i=1}^{N}(R_H(T_i)-R_V(T_i))^2}{N\times(N^2-1)},$$

其中,$R_H(T_i)$和 $R_V(T_i)$分别是机器译文 T_i 自动估计的排名序号和人工评价的排名序号. 斯皮尔曼相关系数越大,译文质量估计与人工评价的相关性越高. 同时也给出了德尔塔平均值(delta average, DeltaAvg)[31]作为排名评价的参考指标.

3 神经网络特征

为了克服译文质量估计中传统特征提取方法严重依赖句子语言学分析等问题,本文结合深度学习

方法,从源语言句子和其机器译文中提取描述翻译质量的特征,提取的特征包括句子向量特征和递归神经网络语言模型特征.

3.1 句子向量特征

3.1.1 词向量训练方法

为了提取句子向量特征,首先需要训练词语的向量(word embedding),使用的词向量训练方法包括3种:

1) 上下文单词预测模型训练词向量方法(word2vec). Mikolov 等人在 Bengio 提出的神经网络语言模型的基础上,去除了比较耗时的隐层,提出了2个简化的神经网络模型用于词向量的训练,分别称为连续的词袋模型(continuous bag-of-words, CBOW)和 Skip-Gram 模型[26]. CBOW 模型是给定上下文单词预测中间单词出现的条件概率,而 Skip-Gram 模型则是根据中间单词预测上下文单词出现的条件概率. 由于 CBOW 模型训练速度快且更适合大规模的数据集,因此实验中使用它训练源语言词语和目标语言词语的词向量. 在词向量训练时,设窗口大小 $window=10$,负采样优化方法中负例的个数 $negative=10$,高频词亚采样频率 $sample=$1e-5,训练迭代次数 $iter=15$.

2) 矩阵分解模型训练词向量方法(Glove). 除了利用上下文单词预测模型训练词向量,我们也尝试了矩阵分解模型 Glove[27] 训练词向量. Glove 基于词语共现关系进行建模,它能有效地结合矩阵分解模型和上下文单词预测模型的优点. 在使用 Glove 模型训练词向量时,将 x_max 参数设为100,窗口大小设为15,训练迭代次数设为50,学习率设为0.75.

3) 连续空间语言模型训练词向量方法(CSLM). Schwenk 在 Bengio 提出的神经网络语言模型的基础上引入多个隐层,利用连续空间语言模型(continuous space language model, CSLM)计算句子语言模型概率和进行词向量训练[32]. Shah 等人在 WMT'16 QE 子任务中利用该模型提取句子的交叉熵和句子向量特征[22]. 为了与 Shah 等人提出的方法进行比较,实验中对于 CSLM 采用了与其一样的参数设置进行词向量训练,即使用4个隐层,投影层使用320个神经元,其他3个隐层每层使用1024个神经元,输出层使用 softmax 激活函数.

3.1.2 句子向量提取策略

获得了词向量之后如何获得句子向量呢?假设词汇表中每一个词 w 的向量表示为 $\boldsymbol{v}_w$,长度为 p 的源语言句子 $S=(s_1,s_2,\cdots,s_p)$ 和其长度为 q 的机器译文 $T=(t_1,t_2,\cdots,t_q)$ 可以使用向量分别表示为

$$\boldsymbol{V}_S=(\boldsymbol{v}_{s_1},\boldsymbol{v}_{s_2},\cdots,\boldsymbol{v}_{s_p}),\ \boldsymbol{V}_T=(\boldsymbol{v}_{t_1},\boldsymbol{v}_{t_2},\cdots,\boldsymbol{v}_{t_q}).$$

为了将源语言句子和机器译文中词向量转化为句子的向量表示,并统一转化后句子向量的维数,我们尝试了4种策略:

1) 算术平均方法(mean). 对于源语言句子或其机器译文,句子向量 $\boldsymbol{V}$ 可以表示为句子中所有词语词向量的算术平均.

$$\boldsymbol{V}_S=\frac{1}{p}\sum_{i=1}^{p}\boldsymbol{v}_{s_i},\ \boldsymbol{V}_T=\frac{1}{q}\sum_{i=1}^{q}\boldsymbol{v}_{t_i}.$$

如果句子中某个词为未登录词,不失一般性,这里将其设为 $\boldsymbol{0}$ 向量.

2) tf-idf 加权平均方法(tf-idf). 由于句子中每一个词对整句的重要性不同,比如在整个语料中出现频率低而在句子中出现频率高的词更能显著表达句子的含义. 为了区分词语的重要性,借鉴于信息检索中 tf-idf 方法对词向量进行加权. 对于源语言句子或其机器译文,其句子向量 $\boldsymbol{V}$ 可以表示为句子中所有词语词向量的 tf-idf 值的加权平均:

$$\boldsymbol{V}_S=\frac{1}{p}\sum_{i=1}^{p}(\text{tf-idf})_{s_i}\times\boldsymbol{v}_{s_i},$$

$$\boldsymbol{V}_T=\frac{1}{q}\sum_{i=1}^{q}(\text{tf-idf})_{t_i}\times\boldsymbol{v}_{t_i}.$$

3) 最小值方法(min). 对于源语言句子或机器译文,句子向量 $\boldsymbol{V}$ 的第 k 维值表示为

$$V_S[k]=\min\ \boldsymbol{v}_w[k],\ w\in\{s_1,s_2,\cdots,s_p\},$$

其中 $k=1,2,\cdots,d$,d 为词向量的维数. 依次类推,最大值方法(max)选择最大值作为最终句子向量的第 k 维.

4) 乘法方法(mul). 对于源语言句子或其机器译文,句子向量的第 k 维表示为句子中所有词语向量的第 k 维连乘的积.

$$V_S[k]=\prod_{i=1}^{p}v_{s_i}[k],\ V_T[k]=\prod_{i=1}^{q}v_{t_i}[k].$$

为了避免句子向量为 $\boldsymbol{0}$ 导致信息丢失,如果句子中出现未登录词,这里将其设为单位向量 $\boldsymbol{1}$.

获取了源语言句子和其机器译文的向量表示 $\boldsymbol{V}_S$ 和 $\boldsymbol{V}_T$ 后,将它们连接成 d_S+d_T 维向量作为译文质量估计的句子向量特征. d_S 和 d_T 分别为源语言词向量和目标语言词向量的维数,由于源语言句子和机器译文在任务中的重要性不同,源语言词语的向量维数和目标语言词语的向量维数不一定相同.

3.2 递归神经网络语言模型特征

由于句子向量特征中,词向量训练方法采用的

是词袋模型,它忽略了机器译文中词序对译文质量的影响.为了刻画机器译文的流利度,进一步引入了源语言句子和其机器译文的递归神经网络语言模型概率作为特征.

传统的统计语言模型在高阶文法概率估计时由于参数空间过大容易导致数据稀疏,递归神经网络语言模型(recurrent neural network language model, RNNLM)通过将词语投影到连续的空间,并在该空间对语言模型进行建模来缓解维数灾难的问题,它已在口语识别任务和统计机器翻译译文重排序任务中实验证明优于传统的统计语言模型[33].因此,我们使用递归神经网络语言模型来计算源语言句子和其机器译文的语言模型概率,并把它们与句子向量特征进行结合.递归神经网络语言模型训练时,它的隐层大小设为 100,后传步数 *bptt* 设为 4,输出层类数设为 200.

4 实 验

4.1 实验数据

为了验证基于神经网络特征的译文质量估计效果,我们在 WMT'15 QE 和 WMT'16 QE 句子级别译文质量估计子任务[28-29]上进行了实验. WMT'15 QE 任务评价英语到西班牙语方向的翻译质量,而 WMT'16 QE 任务评价英语到德语方向的翻译质量.实验中仅使用当年官方公布的语料,其规模统计如表 1 所示,其中神经网络特征训练语料为 WMT 评测方发布用于训练统计机器翻译系统的双语平行语料,这里将其源语言端和目标语言端语料分别用来训练词语的词向量和递归神经网络语言模型.在所有语料使用前均对其进行了符号化(tokenizer)处理[34].

Table 1 The Corpus Statistics

表 1 语料规模统计

Corpus		WMT'15 QE		WMT'16 QE	
		English	Spanish	English	German
Neural Network Feature Training Corpus	Number of Sentences	3.8M	3.8M	4.7M	4.7M
	Vocabulary Size	715.6K	875.5K	927.0K	1 786.3K
	Number of Tokens	102.0M	107.4M	119.5M	114.1M
Training Set	Number of Sentences	11 271	11 271	12 000	12 000
	Vocabulary Size	23.1K	24.5K	9.1K	13.9K
	Number of Tokens	232.7K	251.5K	196.8K	205.6K
Development Set	Number of Sentences	1 000	1 000	1 000	1 000
	Vocabulary Size	5.4K	5.6K	2.9K	3.3K
	Number of Tokens	21.1K	22.7K	17.4K	18.9K
Test Set	Number of Sentences	1 817	1 817	2 000	2 000
	Vocabulary Size	7.8K	8.3K	3.8K	4.6K
	Number of Tokens	37.3K	39.9K	30.7K	33.5K

4.2 实验结果

为了比较不同的特征对译文质量估计的性能影响,实验中统一采用支持向量回归方法建立质量估计模型,性能评价的主要指标分别为 Pearson r 和 Spearman ρ,参考指标为 MAE,RMSE 和 DeltaAvg,其中 Pearson r,Spearman ρ 或 DeltaAvg 值越大,表示性能越好;而 MAE 或 RMSE 值越大,表示性能越差.

首先,实验中将本文提出的上下文单词预测模型和矩阵分解模型提取句子向量特征的方法与连续空间语言模型方法进行了对比,为了与 Shah 等人提出的方法[22]进行比较,固定源语言端和目标语言端词向量维数均为 256,采用算术平均方法求取源语言句子和其机器译文的句子向量.表 2 和表 3 分别给出了不同的句子向量特征在 WMT'15 QE 和 WMT'16 QE 任务上的译文质量估计性能,我们发现使用上下文单词预测模型(word2vec(256))和矩阵分解模型(Glove(256))提取句子向量特征的方法在 Pearson r 和 Spearman ρ 相关性指标上均超过了连续空间语言模型方法(CSLM(256)).而连续空间语言模型方法由于在输出层 softmax 激活函数求条

件概率时只考虑高频词(同文献[22]一致,我们取32K高频词),而这些高频词的数量远小于词汇表中词语的数量,在WMT'15 QE中占训练语料目标端词汇量的1/30,而在WMT'16 QE中仅占训练语料目标端词汇量的1/56,这导致它的性能较低.尽管我们考虑增加高频词数量来提高句子向量特征的质量,但是,随着高频词数量的增加,它的算法复杂度将成指数增加,而系统性能的提升有限.为了简化比较,后续实验中均采用上下文单词预测模型提取句子向量特征.

Table 2 The System Performance with Different Features on WMT'15 QE Tasks

表2 使用不同的特征在WMT'15 QE任务上系统的性能

Feature Sets	Scoring			Ranking	
	Pearson r	MAE	RMSE	Spearman ρ	DeltaAvg
CSLM(256)	0.198	**14.253**	18.722	0.182	3.293
word2vec(256)	0.300	14.495	18.116	0.287	4.780
Glove(256)	**0.329**	14.487	**17.970**	**0.320**	**5.699**
Baseline	0.229	14.733	18.394	0.205	3.659
word2vec	0.332	13.613	17.745	0.311	5.645
word2vec+RNNLM	0.354	13.564	17.602	0.329	5.907
word2vec+RNNLM+Baseline	**0.357**	**13.564**	**17.573**	**0.330**	**5.940**

Table 3 The System Performance with Different Features on WMT'16 QE Tasks

表3 使用不同的特征在WMT'16 QE任务上系统的性能

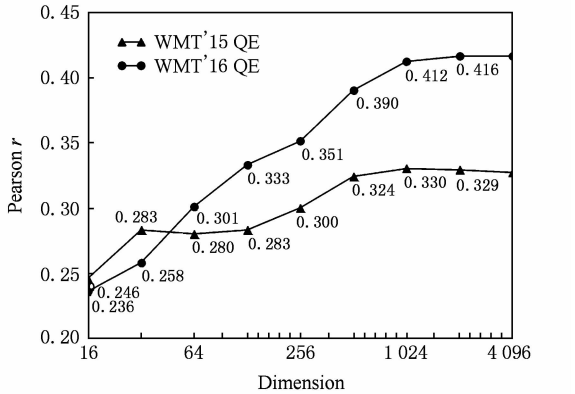
Feature Sets	Scoring			Ranking	
	Pearson r	MAE	RMSE	Spearman ρ	DeltaAvg
CSLM(256)	0.167	14.663	18.960	0.175	3.316
Glove(256)	0.328	13.827	18.474	0.366	5.964
word2vec(256)	**0.351**	**13.642**	**18.209**	**0.391**	**6.388**
Baseline	0.367	13.353	18.055	0.396	6.679
word2vec	0.412	13.160	17.600	0.443	7.445
word2vec+RNNLM	0.441	12.891	17.353	0.468	7.955
word2vec+RNNLM+Baseline	**0.454**	**12.766**	**17.197**	**0.482**	**8.109**

其次,采用最优的词向量维数组合句子向量特征(在后续4.2.1小节和4.2.2小节中讨论),将其与评测方提供的17个基准特征(QuEst方法提取的特征)进行了比较.实验结果表明:单纯采用上下文单词预测模型提取的句子向量特征(word2vec),在WMT'15 QE和WMT'16 QE任务上译文质量估计的效果均显著的优于QuEst基准特征(Baseline)的性能.进一步,将句子向量特征与递归神经网络语言模型特征(RNNLM)结合,在WMT'16 QE任务上打分相关性系数Pearson r由0.412提高到0.441,提高了7.0%,而排名相关性系数Spearman ρ由0.443提高到0.468,提高了5.6%.这说明递归神经网络语言模型特征对提高译文质量估计性能起着很大的作用.最后将Baseline特征与神经网络特征进行融合,系统性能在WMT'15 QE任务上提高不显著,而在WMT'16 QE任务上打分相关性系数Pearson r和排名相关性系数Spearman ρ分别提高了2.9%和3.0%.这些实验对比表明,本文提出的神经网络特征能够较好地描述翻译的质量,使用神经网络特征系统性能较QuEst方法有了显著提高,最高提升达到54.6%(0.229→0.354).

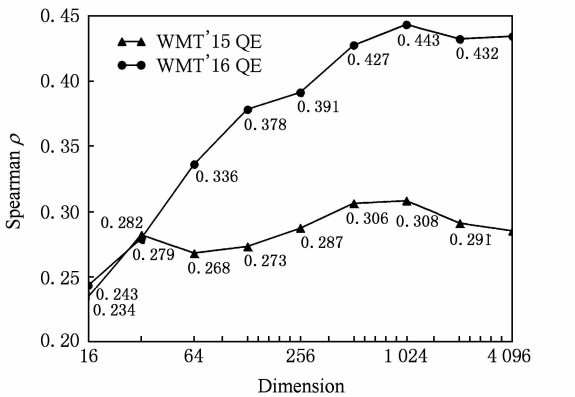
4.2.1 词向量维数对性能的影响

为了揭示词向量维数对译文质量估计性能的影响,实验中将句子向量生成方式固定为算术平均法.首先当源语言词向量维数和目标语言词向量维数相同时,不断增加维数值,实验结果如图1所示,在WMT'15 QE任务中当向量维数为1 024时在打分任务和排序任务都取得了最好的结果,在WMT'16

QE 任务中当向量维数为 2 048 和 1 024 时分别在打分任务和排序任务取得了最好的结果.



(a) Pearson's correlation coefficient

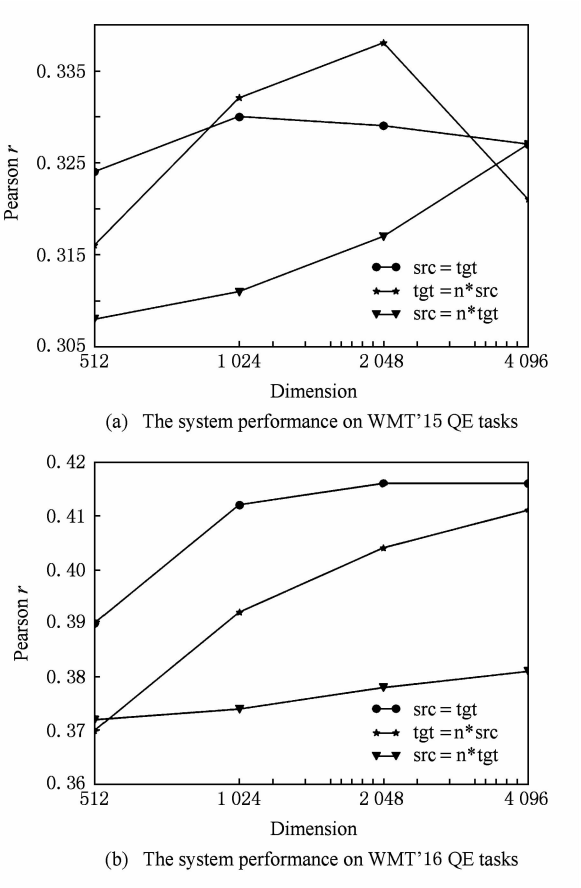


(b) Spearman's correlation coefficient

Fig. 1 Changes in system performance based on the simultaneous growth of word vector dimensions

图 1 词向量维数同步变化对系统的影响

Mikolov 等人实验证明在机器翻译中将源语言向量维数设置为目标语言向量维数的 2～4 倍时，翻译质量最好[35]. 借鉴这个思路，实验中固定源语言(src)或目标语言(tgt)词向量维数为 256 维，让另一端语言维数按 n 倍增长，n 的取值为 2，4，8，16. Pearson 相关系数的变化曲线如图 2 所示，在 WMT'15 QE 任务中当源语言词向量维数为 256 维，目标语言词向量维数为其 8 倍时系统性能最优，而在 WMT'16 QE 任务中，当源语言词向量维数和目标语言词向量维数都为 2 048 维时系统性能最优. 由于在译文质量估计中机器译文特征比源语言句子特征更重要，我们发现增加目标语言词向量维数比增加源语言词向量维数更能提高系统性能. 当然这不能说明源语言特征不重要，实验结果同时表明单独增加源语言词向量维数也能逐步提高系统性能. 这吻合了 Paetzold 等人得出的"源语言句子对于预测目标语言句子的质量有着很大的作用"的结论[23].



(a) The system performance on WMT'15 QE tasks

(b) The system performance on WMT'16 QE tasks

Fig. 2 Changes in system performance based on variation of single word vector dimension

图 2 单一端词向量维数变化对系统的影响

4.2.2 句子向量特征提取策略实验对比

为了比较 3.1 节提出的不同句子向量特征提取策略，实验中将词向量维数固定为 256，分别使用算术平均方法(mean)、tf-idf 加权平均方法(tf-idf)、最小值方法(min)、最大值方法(max)和乘法方法(mul)提取句子特征，它们在 WMT'15 QE 和 WMT'16 QE 任务中的性能如表 4 和表 5 所示，其中采用算术平均方法将句子向量化表示基本取得了最优的相关性，tf-idf 加权平均方法尽管对词向量设置了不同的权重，这些权重对信息检索起着重要作用，但是在译文质量估计中效果不明显.

由于训练词向量和递归神经网络语言模型需要一定规模的单语语料，本文通过实验比较了不同的语料规模对抽取的神经网络特征质量的影响，限于篇幅，这里没有给出结果数据，从实验中发现当训练语料句子规模在 1 M 以上时，系统性能基本没有降低，而当语料规模少于 1 M，随着语料规模的减少，系统性能会逐步降低. 这说明词向量和递归神经网络语言模型训练对语料规模的依赖并不大.

Table 4 The Performance of Different Sentence Embedding Feature Extraction Strategies on WMT'15 QE

表 4 不同句子向量特征提取策略在 WMT'15 QE 中的性能

Extraction Strategy	Scoring			Ranking	
	Pearson r	MAE	RMSE	Spearman ρ	DeltaAvg
mul	−0.001	15.113	18.857	0.020	0.331
min	0.227	14.711	18.433	0.222	3.980
max	0.247	14.799	18.366	0.238	4.165
tf-idf	0.269	**14.414**	18.185	**0.295**	4.753
mean	**0.300**	14.495	**18.116**	0.287	**4.780**

Table 5 The Performance of Different Sentence Embedding Feature Extraction Strategies on WMT'16 QE

表 5 不同句子向量特征提取策略在 WMT'16 QE 中的性能

Extraction Strategy	Scoring			Ranking	
	Pearson r	MAE	RMSE	Spearman ρ	DeltaAvg
mul	0.048	15.115	18.962	0.057	1.180
min	0.238	14.513	18.473	0.253	4.511
max	0.245	14.278	18.482	0.297	5.337
tf-idf	0.248	14.421	18.419	0.269	4.715
mean	**0.351**	**13.642**	**18.209**	**0.391**	**6.388**

5 结束语

本文提出利用神经网络特征,包括句子向量特征和递归神经网络语言模型特征,来提高译文质量估计与人工评价的相关性,并通过实验验证本文方法优于传统的 QuEst 方法和基于连续空间语言模型的特征提取方法. 与译文质量估计中基于语言学分析提取特征的方法相比,利用神经网络提取特征不仅提高了译文质量估计的性能,而且方法与语言种类无关;它的缺点在于提取的特征解释性不强,且词向量和语言模型训练时需要相关语言的单语语料,幸运的是随着互联网的发展,网络上存在大量的单语语料可供使用. 在以后的工作中,我们将探索将神经网络应用到译文质量估计模型构建中,创建一个端到端的系统.

参 考 文 献

[1] Gandrabur S, Foster G. Confidence estimation for translation prediction [C] //Proc of the 7th Conf on Natural Language Learning at HLT-NAACL. Stroudsburg, PA: ACL, 2003: 95-102

[2] Ueffing N, Ney H. Word-level confidence estimation for machine translation [J]. Computational Linguistics, 2007, 33(1): 9-40

[3] Blatz J, Fitzgerald E, Foster G, et al. Confidence estimation for machine translation [C] //Proc of the 20th Int Conf on Computational Linguistics. Stroudsburg, PA: ACL, 2004: 315-321

[4] Quirk C. Training a sentence-level machine translation confidence measure [C] //Proc of the 4th LREC. Paris: ELRA, 2004: 825-828

[5] Ning Wei, Miao Xuelei, Hu Yonghua, et al. Machine translation quality evaluation without reference based on SVM [C] //Proc of Machine Trans Research Progress—The 4th National Conf on Machine Translation. Beijing: Chinese Information Processing Society of China, 2008: 196-203 (in Chinese)

(宁伟, 苗雪雷, 胡永华, 等. 基于 SVM 的无参考译文的译文质量评测[C] //机器翻译研究进展——第四届全国机器翻译研讨会论文集. 北京: 中国中文信息学会, 2008: 196-203)

[6] Yin Baosheng, Miao Xuelei, Ji Duo, et al. Research on automatic translation quality evaluation technology without translation references for large-scale translations [J]. Journal of Shenyang Aerospace University, 2012, 29(1): 70-74 (in Chinese)

(尹宝生, 苗雪雷, 季铎, 等. 大规模无参考译文质量自动评测技术的研究[J]. 沈阳航空航天大学学报, 2012, 29(1): 70-74)

[7] Specia L, Shah K, De Souza J G C, et al. QuEst-A translation quality estimation framework [C] //Proc of ACL: System Demonstrations. Stroudsburg, PA: ACL, 2013: 79-84

[8] Rubino R, Toral A, Vaíllo S C, et al. The CNGL-DCU-Prompsit translation systems for WMT13 [C] //Proc of the 8th WMT. Stroudsburg, PA: ACL, 2013: 211-216

[9] Soricut R, Bach N, Wang Z. The SDL language weaver systems in the WMT12 quality estimation shared task [C] //Proc of the 7th WMT. Stroudsburg, PA: ACL, 2012: 145-151

[10] Hardmeier C, Nivre J, Tiedemann J. Tree kernels for machine translation quality estimation [C] //Proc of the 7th WMT. Stroudsburg, PA: ACL, 2012: 109-113

[11] Almaghout H, Specia L. A CCG-based quality estimation metric for statistical machine translation [C] //Proc of the XIV MT Summit. Langhorne, PA: AMTA, 2013: 223-230

[12] Avramidis E. Quality estimation for machine translation output using linguistic analysis and decoding features [C] //Proc of the 7th WMT. Stroudsburg, PA: ACL, 2012: 84-90

[13] Bengio Y, Ducharme R, Vincent P, et al. A neural probabilistic language model [J]. Journal of Machine Learning Research, 2003, 3(2): 1137-1155

[14] Cho K, Bahdanau D, Bougares F, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation [C] //Proc of 2014 Conf on EMNLP. Stroudsburg, PA: ACL, 2014: 1724-1734

[15] Liu Zhiyuan, Sun Maosong, Lin Yankai, et al. Knowledge representation learning: A review [J]. Journal of Computer Research and Development, 2016, 53(2): 247-261 (in Chinese)
(刘知远, 孙茂松, 林衍凯, 等. 知识表示学习研究进展[J]. 计算机研究与发展, 2016, 53(2): 247-261)

[16] Shah K, Logacheva V, Paetzold G, et al. SHEF-NN: Translation quality estimation with neural networks [C] // Proc of the 10th WMT. Stroudsburg, PA: ACL, 2015: 342-347

[17] Kreutzer J, Schamoni S, Riezler S. Quality estimation from ScraTCH (QUETCH): Deep learning for word-level translation quality estimation [C] //Proc of the 10th WMT. Stroudsburg, PA: ACL, 2015: 316-322

[18] Patel R N, Sasikumar M. Translation quality estimation using recurrent neural network [C] //Proc of the 1st Conf on Machine Translation. Stroudsburg, PA: ACL, 2016: 819-824

[19] Scarton C, Beck D, Shah K, et al. Word embeddings and discourse information for machine translation quality estimation [C] //Proc of the 1st Conf on Machine Translation. Stroudsburg, PA: ACL, 2016: 831-837

[20] Schwenk H. Continuous space language models [J]. Computer Speech & Language, 2007, 21(3): 492-518

[21] Shah K, Ng R W M, Bougares F, et al. Investigating continuous space language models for machine translation quality estimation [C] //Proc of EMNLP 2015. Stroudsburg, PA: ACL, 2015: 1073-1078

[22] Shah K, Bougares F, Barrault L, et al. SHEF-LIUM-NN: Sentence level quality estimation with neural network features [C] //Proc of the 1st Conf on Machine Translation. Stroudsburg, PA: ACL, 2016: 838-842

[23] Paetzold G H, Specia L. SimpleNets: Machine translation quality estimation with resource-light neural networks [C] // Proc of the 1st Conf on Machine Translation. Stroudsburg, PA: ACL, 2016: 812-818

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

[25] Kim H, Lee J. A recurrent neural networks approach for estimating the quality of machine translation output [C] // Proc of NAACL-HLT 2016. Stroudsburg, PA: ACL, 2016: 494-498

[26] Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space [J]. Proceedings of ICLR, arXiv: 1301.3781, 2013

[27] Pennington J, Socher R, Manning C D. Glove: Global vectors for word representation [C] //Proc of EMNLP 2014. Stroudsburg, PA: ACL, 2014: 1532-1543

[28] Bojar O, Chatterjee R, Federmann C, et al. Findings of the 2015 workshop on statistical machine translation [C] //Proc of the 10th WMT. Stroudsburg, PA: ACL, 2015: 1-46

[29] Bojar O, Chatterjee R, Federmann C, et al. Findings of the 2016 conference on machine translation [C] //Proc of the 1st Conf on Machine Translation. Stroudsburg, PA: ACL, 2016: 131-198

[30] Snover M, Dorr B, Schwartz R, et al. A study of translation edit rate with targeted human annotation [C] //Proc of AMTA 2006. Langhorne, PA: AMTA, 2006: 223-231

[31] Callison-Burch C, Koehn P, Monz C, et al. Findings of the 2012 workshop on statistical machine translation [C] //Proc of the 6th WMT. Stroudsburg, PA: ACL, 2011: 10-51

[32] Schwenk H. Continuous space translation models for phrase-based statistical machine translation [C] //Proc of COLING 2012. New York: ACM, 2012: 1071-1080

[33] Mikolov T, Karafiát M, Burget L, et al. Recurrent neural network based language model [C] //Proc of Interspeech 2010. Grenoble, France: ISCA, 2010: 1045-1048

[34] Koehn P, Hoang H, Birch A, et al. Moses: Open source toolkit for statistical machine translation [C] //Proc of the 45th Annual Conf on ACL. Stroudsburg, PA: ACL, 2007: 177-180

[35] Mikolov T, Le Q V, Sutskever I. Exploiting similarities among languages for machine translation [J]. arXiv preprint arXiv: 1309.4168, 2013

Chen Zhiming, born in 1993. Postgraduate. His main research interests include natural language processing and machine translation.

Li Maoxi, born in 1977. PhD, associate professor. Member of CCF. His main research interests include natural language processing and machine translation.

Wang Mingwen, born in 1964. PhD, professor and PhD supervisor. Senior Member of CCF. His main research interests include natural language processing and information retrieval.