文章编号：

基于自编码器的通用性文本表征*

张明华 [1]，吴云芳 [1]，李伟康 [1]，张仰森 [2]

（1 北京大学计算语言学教育部重点实验室，北京，100871；

2 北京信息科技大学计算机学院，北京，100871）

摘要：为了学习文本的语义表征，之前的研究者主要依赖于复杂的循环神经网络（recurrent neural networks, RNNs）和监督式学习方法。本文中，我们提出了一种门控联合池化自编码器（gated mean-max AAE），用于学习通用性的文本语义表征。本文的自编码器完全通过多头自注意力机制（multi-head self-attention mechanism）来构建编码器和解码器网络。在编码阶段，提出了均值−最大化（mean-max）联合表征策略，即同时运用平均池化（mean pooling）和最大池化（max pooling）操作来捕获输入文本中更多样性的语义信息。为促使联合池化表征可以全面地指导重构过程，解码器采用门控操作进行动态关注。通过在大规模中英文未标注语料上训练模型，我们获得了高质量的句子编码器。在重构文本段落的实验中，本文模型在实验效果和计算效率上均超越了传统的 RNNs 模型。我们将公开训练好的文本编码器，使其可以方便地运用于后续的研究。

关键词：文本表征；自编码器；多头自注意力机制

中图分类号：TP391　　　　　**文献标识码：**A

Gated Mean-Max Autoencoder for Paragraph Representations

Minghua Zhang[1], Yunfang Wu[1], Weikang Li[1], Yangsen Zhang[2]

(1. MOE Key Laboratory of Computational Linguistics, Peking University, Beijing, 100871, China;

2. Computer School, Beijing Information Science and Technology University, Beijing, 100871, China)

Abstract: In order to learn distributed representations of text sequences, the previous methods focus on complex recurrent neural networks or supervised learning. In this paper, we propose a gated mean-max autoencoder both for Chinese and English text representations. In our model, we rely entirely on the multi-head self-attention mechanism to construct the encoder and decoder. In the encoding we propose a mean-max strategy that applies both mean and max pooling operations over the hidden vectors to capture diverse information of the input. To enable the information to steer the reconstruction process, the decoder employ elementwise gate to select between mean and max representations dynamically. By training our model on a large amount of Chinese and English unlabelled data separately, we obtain high-quality text encoders which we will make publicly available. Experimental results of reconstructing coherent long texts from the encoded representations demonstrate the superiority of our model over the traditional recurrent neural network, both in the performance quality and computational complexity.

Key words: text representations; autoencoder; multi-head self-attention

1　引言

自动学习文本（词，短语，句子以及段落）的语义表征是自然语言处理（natural language processing, NLP）中一个基础性的研究工作。目前已经存在非常高效的模型[1]利用大量无标语料来学习词表征（word embedding），这些词表征为 NLP 的各种下游任务提供了有效的

* **收稿日期：**　　　　　　**定稿日期：**

基金项目：国家自然科学基金项目（61773026, 61572245）

作者简介：

特征支持。近年来，越来越多的研究者开始关注更大文本单元的语义表征，其目标在于把文本序列中包含的语法和语义信息编码成一个固定长度的向量，然后将这些学习到的表征知识迁移到其他的 NLP 任务中。

其中一个被广泛研究的方向是使用基于 RNNs 的编码器-解码器架构[2,3,4,5]，在给定输入文本序列的前提下，以重构输入序列或者预测上下文序列为训练目标来学习文本的语义表征。另外一些研究者[6,7]提出借助标注数据来学习一个通用的语义编码器，比如斯坦福的自然语言推理数据集（Stanford Natural Language Inference, SNLI）[8]。然而这些现存的方法均有其固有的限制。首先，由于序列化的建模方式，RNNs 网络在面对长文本单元（段落）的时候，显得极为耗时，尤其是在需要大规模训练数据来学习文本语义表征的场景下，训练开销变得更加难以接受。比如说，为了获取有意义的向量表征，[2]花费了两周时间来训练 skip-thought vector。其次，对于 SNLI 这种大规模的高质量标注数据，在其他语言中基本上是不存在的，比如本文所处理的中文文本。

本文的研究致力于以无监督的方式来自动学习文本语义表征，我们提出了一种门控联合池化自编码器（gated mean-max AAE）。具体地，对于输入的文本序列，编码器网络首先执行多头自注意力操作以获取输入的隐向量；然后联合使用平均池化和最大池化产出 mean-max 语义表征。在重构过程中，解码器先利用多头自注意力操作关注文本序列之前时间步的词，接着运用门控机制来动态关注 mean-max 表征。因此，在重构的每一步，解码器不仅仅依赖之前时间步的信息，而且会充分地利用输入文本的整体语义信息。另一方面，由于联合池化操作的使用，解码器可以得到不同语义表征空间的指导，以满足不同文本单元的重构需求。

针对不同的语言，我们分别在公开的英文数据集[9]和大规模的中文 Gigaword 语料（LDC2005T14）上训练文本编码器。为了验证向量表征捕捉文本序列语义信息的能力，我们从 mean-max 表征出发来重构中英文的长文本段落。实验结果显示，在英文数据集上，我们的模型超越了基于注意力的层次化 LSTM（Long- short term memory）网络。在中文数据集上，本文模型的实验效果也远超传统 RNNs 模型。同时，我们模型的并行建模方式提高了训练效率，在相同数据量的情况下，相比于 RNNs 模型 71 小时的训练时间，我们的模型仅用了 32 小时。为了促进相关研究，我们将公开已经训练好的文本编码器，对于输入的一个文本序列，输出定长的分布式向量表示。

本文的主要贡献在于：

（1） 将多头自注意力机制引入自编码器，用于学习通用性的文本语义表征。由于并行的建模方式，在面临大规模无标数据时，极大缩减了训练时间，可以更高效的获取有意义的向量表征。

（2） 同时运用平均池化和最大池化操作来获取文本序列的联合语义表征，然后通过门控机制，让两个不同的表征空间共同动态地指导解码过程。

（3） 首次以无监督的方式训练高质量的中文通用性文本编码器，并公开发布，以促进中文信息处理的相关研究。

2 相关工作

近年来，随着深度学习技术在自然语言处理领域的迅猛发展，越来越多的研究者开始关注文本序列的语义表征，并且提出了各种各样的模型来尝试解决这个问题。相关研究可以划分为两个方向。

由于互联网中海量生语料的存在，直接从大规模未标注数据中学习语义表征已经成为了一个热门的研究方向。[10]提出了 paragraph2vec 模型，通过在对数线性神经语言模型中[11]引入全局的段落向量来学习文本的分布式表征。在 skip-thoughts 模型[2]中，作者提出使用一个 RNN 网络来编码输入的句子，使用另外两个 RNN 网络分别预测文档中的上一个句子

和下一个句子，但是模型训练较为耗时。[3]通过在 skip-thoughts 模型中引入层正则化操作（layer normalization）提高了模型的训练速度，在迁移任务上也取得了更好的效果，但是文中的模型仍然花费了一个月的训练时间。[4]提出了序列降噪自编码器（sequential denoising autoencoder, SDAE），从含有噪音的输入序列中重构原始的句子。另外作者也实现了一些简单的词袋模型，比如 word2vec-SkipGram 和 word2vec-CBOW。[12]提出了一个简单高效的平滑逆频方法（Smooth Inverse Frequency, SIF），通过词向量的加权平均来计算句子的向量表征。[5]使用卷积神经网络（convolutional neural network, CNN）编码器和 LSTM 解码器来重构输入句子和预测上下文序列，同时作者探索了利用层次化的模型来预测文档中连续的多个下文句子。

相比而言，高质量的人工标注数据就显得较为稀缺，但是研究者对各种监督方法依然进行了有效的探索。[4]尝试了三种不同的标注资源，包括使用神经语言模型将词语的字典定义映射到相应的词向量，将图像字幕映射到图像向量，以及用双语平行语料训练神经机器翻译模型，来学习通用的语义编码器。[6]认为自然语言推理任务涉及丰富的句子间语义关系，因此他们提出使用 SNLI 数据集来学习句子表征，并且通过对比 7 种不同的模型架构，他们发现基于最大池化操作的双向 LSTM 网络取得了最好的实验效果。[7]提出利用多任务学习机制来训练句子编码器，他们使用的任务包括 skip-thought 模型[2]中的预测上下句任务，自然语言推理任务（SNLI），以及对话回复任务[13]，作者还探索了结合句子级别的语义和词级别的信息用于迁移任务。

3 多头自注意力机制

注意力机制最早由[15]提出并成功地用于神经网络机器翻译中，其主要思想可以描述为：给定一系列的键值对(K,V)和查询向量q，首先在q和K的每一个键之间计算归一化的权重，然后对相应的值向量V进行加权求和，以生成动态的注意力向量，使得和当前查询向量相关性更高的键值对能够贡献更多的语义信息。而在自注意力操作中，查询向量和键值对均来自相同的输入序列。本文使用的自注意力运用多头机制[14]，会并行地多次调用注意力操作，然后将每个注意力操作的结果串接起来。具体而言，对于q和(K,V)，通过下面的公式计算注意力向量a：

$$a = self\_attention(q,K,V) \tag{1}$$

$$= concat(head_1,\cdots,head_l) \tag{2}$$

$$head_i = attention(\bar{q},\bar{K},\bar{V}) \tag{3}$$

$$= softmax\left(\frac{\bar{q}\bar{K}^T}{\sqrt{d_k}}\right)\bar{V} \tag{4}$$

其中：

$$\bar{q},\bar{K},\bar{V} = qW_i^q, KW_i^K, VW_i^V \tag{5}$$

W_i^q, W_i^K和W_i^V均为参数矩阵；$\bar{q} \in \mathbb{R}^{d_k}, \bar{K} \in \mathbb{R}^{n_k \times d_k}, \bar{V} \in \mathbb{R}^{n_k \times d_v}$；$d_k$和$d_v$分别是$\bar{K}$和$\bar{V}$的维度；$n_k$是$(K,V)$中包含的键值对数量。

多头自注意力机制允许模型动态地关注不同键值对的信息，可以适应不同查询的信息需求。而且由于每个$head$计算维度的减小以及多头的并行操作，所以其总的计算开销和普通的单头注意力操作相当。

4 模型

我们的模型采用编码器-解码器结构，如图 1 所示。输入序列$\mathbf{x}$首先通过左边的编码器网络（encoder）转换为隐向量序列，然后通过中间衔接部分的平均池化和最大池化操作得到输入的语义表征，最后右边的解码器网络（decoder）利用语义表征来重构输入序列。

4.1 编码器

在建模过程中，本文统一将输入的文本视为词序列，不区分句子或段落。此处使用$S=\{w_1,w_2,\cdots,w_N\}$来表示模型的输入词序列，其中N代表输入序列中词的数量。同时在每个输

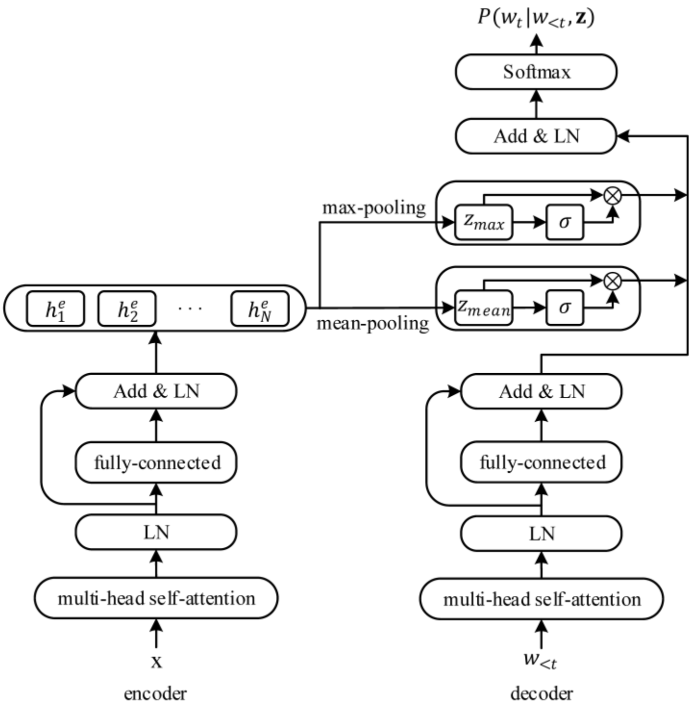

图 1 门控 mean-max 自编码器网络结构

入序列的末尾添加序列结束符“$</S>$”。序列S中的每一个词w_t首先均被转换为k维的词向量$e_t=W_e[w_t]$,其中$W_e\in\mathbb{R}^{d_w\times V}$是词向量矩阵，$V$是模型的词表，$w_t$表示$S$中的第$t$个词在词表中的索引，$W_e[v]$则表示矩阵$W_e$的第$v$列。

另外为了建模输入序列的序信息，我们对输入的每一个词都增加位置编码[14]：

$$p_t[2i]=sin\left(t/10000^{2i/d_w}\right) \tag{6}$$
$$p_t[2i+1]=cos\left(t/10000^{2i/d_w}\right) \tag{7}$$

其中t表示输入序列的第t个词位，i是位置编码的第i维。位置编码中的偶数维对应一个正弦值，奇数维则对应一个余弦值。由此，可以得到编码器的输入为$x_t=e_t+p_t$。

编码器网络包含一个多头自注意力层（multi-head self-attention layer）和一个全连接层（fully connected layer），负责将输入的词序列转换为隐向量序列。有别于[14]的建模过程，我们移除了多头自注意力层的残差连接（residual connection），仅保留全连接层的残差连接，这样我们的模型可以自由地扩展隐层的维度，以使隐向量包含更多的语义信息。

给定输入$\mathbf{x}=(x_1,\cdots,x_N)$,编码器第$t$步的隐向量$h_t^e$可以通过下面的公式计算得到：

$$a_t^e=self\_attention(x_t,\mathbf{x},\mathbf{x}) \tag{8}$$
$$\bar{a}_t^e=LN(a_t^e) \tag{9}$$
$$f_t^e=max(0,\ \bar{a}_t^e W_1^e+b_1^e)W_2^e+b_2^e \tag{10}$$
$$h_t^e=LN(f_t^e+\bar{a}_t^e) \tag{11}$$

其中$W_1^e\in\mathbb{R}^{d_m\times d_f}$和$W_2^e\in\mathbb{R}^{d_f\times d_m}$均为参数矩阵；$b_1^e\in\mathbb{R}^{d_f}$和$b_2^e\in\mathbb{R}^{d_m}$均为偏置向量；$d_m$和$d_f$分别是隐向量和全连接子层的维度；$LN$表示层正则化操作。

由于采用多头自注意力机制作为网络的基本构建块，我们的模型可以非常有效地进行并行操作。对于任意长度的输入序列，模型可以同时输出所有的隐向量，相比于 LSTM 模块的序列处理方式，这将极大地减小计算复杂度。

4.2 语义表征

给定任意输入的隐向量序列$\{h_t^e\}_{t=[1,\cdots,N]}$，需要将这一系列的局部隐向量表征转换为一个全局的语义表征。我们选择池化操作，这不仅使得提取的语义表征独立于输入序列的长度，同时可以获得固定长度的向量表征。[6]在 BiLSTM 网络的基础上，通过池化操作来获取输入句子的语义表征，他们实验了平均池化和最大池化，结果表明最大池化操作获得的句子表征在各种迁移任务上表现得更好。

而在本文中，我们提出同时运用平均池化和最大池化操作。对于给定的向量序列，最大池化计算各维度的最大值，以试图捕获序列中最显著的属性而过滤掉其他含有较少信息量的局部值。而平均池化不筛选显著的局部值，也不关注序列中那些特殊的特征值，而是捕捉更普适性的信息。很显然，这两种池化策略可以互为补充。通过并联两种池化操作的结果，得到输入序列的 mean-max 表征：

$$z_{max}[i] = \max_i h_{ti}^e \tag{12}$$

$$z_{mean} = \frac{1}{N}\sum_t h_t^e \tag{13}$$

$$\mathbf{z} = [z_{max}, z_{mean}] \tag{14}$$

通过联合使用两种不同的池化策略，可以从不同的角度来处理隐向量序列，以捕捉更多样性的语义特征，这使 mean-max 表征可以为重构过程提供更多的信息，在面对长文本序列的时候展现出强大的鲁棒性。

4.3 解码器

正如编码器网络那样，解码器也运用多头自注意力机制来重构输入序列。如图 1 所示，编码器和解码器通过 mean-max 门控层进行连接，这将使解码器在重构过程中动态地利用 mean-max 信息。

为了促使隐向量的维度扩展，我们只在全连接层和 mean-max 门控层使用残差连接，而去掉多头自注意力层的残差连接。给定之前时间步的词$\mathbf{y} = (x_1, \cdots, x_{t-1})$和输入序列的隐表征$\mathbf{z}$作为解码器的输入，我们可以通过下面的公式计算解码器第t步的隐向量h_t^d：

$$a_t^d = self\_attention(y_t, \mathbf{y}, \mathbf{y}) \tag{15}$$

$$\bar{a}_t^d = LN(a_t^d) \tag{16}$$

$$f_t^d = max(0,\ \bar{a}_t^d W_1^d + b_1^d)W_2^d + b_2^d \tag{17}$$

$$\bar{f}_t^d = LN(f_t^d + \bar{a}_t^d) \tag{18}$$

$$z_{max}^d = z_{max} \otimes \boldsymbol{\sigma}(z_{max} W_3^d + \bar{f}_t^d W_4^d + b_3^d) \tag{19}$$

$$z_{mean}^d = z_{mean} \otimes \boldsymbol{\sigma}(z_{mean} W_5^d + \bar{f}_t^d W_6^d + b_4^d) \tag{20}$$

$$h_t^d = LN(\bar{f}_t^d + z_{max}^d + z_{mean}^d) \tag{21}$$

其中$W_1^d \in \mathbb{R}^{d_m \times d_f}$,$W_2^d \in \mathbb{R}^{d_f \times d_m}$,以及$W_3^d, W_4^d, W_5^d, W_6^d \in \mathbb{R}^{d_m \times d_m}$均为参数矩阵；$b_1^d \in \mathbb{R}^{d_f}$和$b_2^d, b_3^d, b_4^d \in \mathbb{R}^{d_m}$均为偏置向量。$\boldsymbol{\sigma}$表示 sigmoid 函数。

给定解码器的隐向量序列$(h_1^d, \cdots, h_N^d)$，重构出输入序列S的概率计算如下：

$$P(w_t | w_{<t}, \mathbf{z}) \propto exp(h_t^d W_7^d + b_5^d) \tag{22}$$

$$\mathcal{L}(\theta) = \sum_t logP(w_t | w_{<t}, \mathbf{z}) \tag{23}$$

其中$W_7^d \in \mathbb{R}^{d_m \times V}, b_5^d \in \mathbb{R}^V$；模型通过优化目标函数来学习重构输入序列。

5 实验

5.1 实验数据

本文采用了两个大规模的中英文数据集（中文 gigaword 和英文 hotel reviews）来分别学习两种语言的文本语义表征，并通过重构长文本段落来验证其捕获语义信息的能力。

我们从中文 Gigaword 语料的新华日报部分构建了中文文本段落数据集。所有数据采用斯坦福的 CoreNLP 工具包[16]进行分词预处理。词表大小为 33090，其他低频词均用“<UNK>”表示。仅保留词数在 10 到 200 之间，并且未登录词比率小于 2%的文本。经过预处理后，总共获得 4194066 个段落。随机划分之后，训练集、开发集和测试集分别包含的段落数量为：3890004、50000 和 50000。

在 Hotel reviews 数据集[9]中，每条评论包含的单词数量在 50 到 250 之间。作者公开的词表中包含语料中前 25000 个高频的单词。所有数据的未登录词比率均小于 2%，评论的平均单词数量为 124.8。训练集和测试集分别包含 340000 条和 40000 条评论。

5.2 实验设置

本文提出的模型主要包含三个组件：编码器和解码器网络构建块，动态门控机制以及 mean-max 联合池化策略。我们分别实现了相应的基准模型进行对比分析。

首先，编码器和解码器网络均采用多头自注意力机制来构建。为了验证多头自注意力机制的作用，实现了基于 LSTM 的模型：编码器采用双向 LSTM 网络计算隐向量，然后运用联合池化策略获取输入的语义表征，并且串接前向 LSTM 和后向 LSTM 的末尾隐向量用于初始化解码器 LSTM 的状态，每一步解码过程均使用门控操作来关注全局的语义表征。下文中我们将以 gated mean-max RAE 来表示。

另外，在 gated mean-max RAE 的基础上删除门控机制，得到了更简单的基准模型 RAE，以探究在没有语义表征动态指导的情况下解码器的重构效果。

最后，为了证明联合池化策略的优势，在只运用平均池化或者最大池化操作的情况下，分别实现了 Gated mean AAE 和 Gated max AAE，两个模型的其他部分和本文模型保持一致。

本文使用 Adam 算法[17]优化模型目标。词向量使用 Xavier 方法[18]进行随机初始化，并和模型参数一起更新。解码过程采用贪心算法（greedy），最大长度不超过输入文本的 1.5 倍。中英文实验所采用的超参数如表 1 所示。

表 1 实验的超参数设置

超参数	hotel reviews	新华日报
the dimension of word embedding	512	512
the dimension of hidden state	1024	1024
head count	8	8
gradient clipping	5.0	5.0
dropout rate	0.2	0.2
learning rate	0.0002	0.0002
batch size	32	32

所有模型均使用 Tensorflow 实现，并在英伟达 GeForce GTX 1080 GPU 上运行。

5.3 实验结果与分析

实验结果如表 2 所示，采用了两个标准的评价指标：BLEU 和 ROUGE。

在中英文数据集上，本文提出的门控和联合池化策略均取得了最优的效果。在 hotel reviews 数据上，相比了 Attention Hierarchical 模型，Gated mean-max AAE 模型显著地提升了实验效果，将 BLEU 值从 28.5 提高到了 63。另外，Attention Hierarchical 模型的编码器和解码器均由 4 层的 LSTM 网络组成，而本文模型只采用了一层的自注意力和全连接网络。因此，Gated mean-max AAE 实现更加简单，运行也更加高效。

门控机制使语义表征可以动态地指导解码器的重构过程，对比 Gated mean-max RAE 和 RAE 可知，其显著地增强了模型性能，在两个数据集上，BLEU 值分别提升 25.6 和 27.7。而且相比于单一的池化操作，联合池化让文本表征携带更丰富的语义信息，在面临中英文的长文本序列时，也展现了极佳的效果。经分析，编码器和解码器的网络结构对实验效果没有显著的影响，但是自注意力机制可以加速模型训练。另外，Gated mean-max RAE 除了使用门控操作和联合表征，还借助了双向 LSTM 编码器的末尾隐向量来初始化解码器的状态，而本文并未使用末尾隐状态的信息。

多头自注意力机制使用并行的建模方式，对于输入的文本，可以在常数的时间内连接序列中所有位置，而循环神经网络需要的计算开销随输入的长度成线性增长。因此，本文模型极大减小了计算复杂度，同时更容易建立文本中的长距离依赖。在中文编码器的训练过程中，Gated mean-max AAE 在一块 GPU 上训练了 32 个小时，然而 Gated mean-max RAE 使用了 71 个小时，并且占了两块 GPU。

表 2 中英文段落重构实验。其中，“Standard”, “Hierarchical”以及“Attention Hierarchical”均为[9]中的模型。

数据集	模型	BLEU	ROUGE-1	ROUGE-2
Hotel Reviews	Standard	24.1	57.1	30.2
	Hierarchical	26.7	59.0	33.0
	Attention Hierarchical	28.5	62.4	35.5
	Gated mean AAE	40.2	89.4	48.9
	Gated max AAE	40.5	89.6	49.6
	Gated mean-max AAE	63.0	92.3	69.8
	Gated mean-max RAE	61.5	86.6	70.8
	RAE	35.9	67.7	41.9
新华日报	Gated mean AAE	45.7	87.6	59.9
	Gated max AAE	45.8	86.1	61.8
	Gated mean-max AAE	61.8	90.7	74.0
	Gated mean-max RAE	65.3	89.5	77.7
	RAE	37.6	73.1	51.2

动态门控机制让解码器即使在面临长文本的时候，依然可以充分的利用输入序列的全局语义表征。为了说明本文模型对长文本处理的优势，图 2 给出测试集的 BLEU 值随文本长度变化的曲线。随着文本长度的增加，传统 RNNs 模型的性能快速下降，联合表征和门控机制的引入使模型依旧具有不错的表现。表 3 中，我们从中文测试集中随机抽出了两个长文本段落的重构实例。

5.4 句子相似度测试

表 4 中展示了 5 个基于句子相似性的检索实例。所有句子均使用中文 gigaword 语料上训练好的汉语文本编码器进行编码。采用句子语义编码的余弦相似度作为检索的衡量指标。分析检索结果可知，编码器从大规模的无标注数据中能够较为精确地捕捉文本的语义信息。

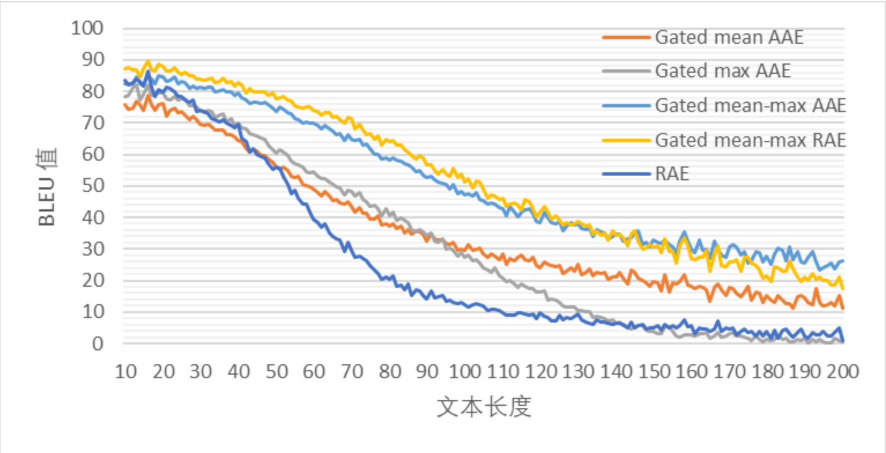

图 2 中文测试集上 BLEU 值随输入长度变化的曲线

表 3 新华日报段落重构实例

Ground Truth	随着《红军长征》卷等一批纪念长征胜利 60 周年的图书的出版发行,解放军历史资料丛书目前已出版了 10 卷、80 多册、6000 多万字,为全军指战员提供了<UNK>特别是<UNK>优良传统教育的生动教材,为促进军队精神文明建设作出了重要贡献。
Gated mean-max AAE	随着《红军长征》卷等一批纪念长征胜利 60 周年的图书的出版发行,解放军历史资料丛书目前已出版了 10 卷、80 多册、丛书,为全军字提供了力所能及<UNK>特别是<UNK>优良传统教育的生动教材,为促进军队精神文明建设作出了重要贡献。
Gated mean-max RAE	随着《红军长征》卷等一批纪念长征胜利 60 周年的图书发行,解放军历史资料出版了丛书要出版几卷、80 多册、字知识,为<UNK>特别是全军长征提供了<UNK>特别是<UNK>优良传统教育的生动教材,为促进军队精神文明建设作出了重要贡献。
RAE	随着《红军长征》卷等一批纪念长征胜利的 60 周年图书出版的发行,有国家图片出版理论技术了北京底的一百个专题组成,2 万字上、全国共有全军生命<UNK>进行了教育胜利及教育比较生动的,培养新干部特别行政区政府起了<UNK>实事求是。
Ground Truth	(记者王艳红)法国医学专家最近提出,英国的新型克雅氏症传染规模可能很小,染病的总人数估计在 200 人左右。
Gated mean-max AAE	(记者尹鸿祝)法国医学专家最近提出,英国的新型的传染规模可能很小,估计的总人数估计在 200 人左右。
Gated mean-max RAE	(记者王艳红)法国医学专家最近提出,英国的新型医学传染规模可能很小,英国的总人数估计在 200 人左右。
RAE	(记者王艳红)法国医学专家最近提出,英国的新型传播香港可能数量很大,向该人数的估计总在 200 人左右。

表 4 中文句子检索。其中第一句是查询输入，第二句是从近 4 百万的新华日报语料中检索出的最相似句子。

中国与东盟关于非传统安全领域合作联合宣言 《中华人民共和国与东盟国家领导人联合宣言》
本届锦标赛将于本月 30 日结束。 本届比赛将于本月 31 日结束。
北京丽人美容有限公司是一家香港独资公司。 允许香港顾问工程公司在内地设立独资公司。
据悉，江苏省政府已作出决定，迅速在全省推广这项科技成果。 据悉，上海市商业部门将在更大的范围内推行这种营销方式。
和平统一、 民族强盛，是全体中国人民的共同愿望。 和平崛起中的中国，期待世界的支持与祝福。

6 结语

本文中，我们提出了门控联合池化自编码器，以从大规模无标注语料中学习不同语言的语义表征。为了提高表征学习效率以及捕捉长距离依赖，使用多头自注意力机制来构建编码器和解码器网络。在编码阶段，平均池化和最大池化操作的联合运用使文本表征可以捕捉更丰富的语义信息。随后，解码器执行门控操作来动态关注编码的全局语义表征。即使处理长文本序列，语义表征依然可以持续有效地指导重构过程。通过语义表征重构中英文的长文本实验中，本文模型的实验效果远优于传统的 RNNs 模型。同时，由于并行化的建模方式，模型极大地缩减了文本表征学习的时间。在后续的研究中，我们将尝试自动从互联网海量数据中构建监督语料，用于学习不同语言的语义表征。

参考文献

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

[2] Kiros R, Zhu Y, Salakhutdinov R R, et al. Skip-thought vectors[C]//Advances in neural information processing systems. 2015: 3294-3302.

[3] Ba J L, Kiros J R, Hinton G E. Layer normalization[J]. arXiv preprint arXiv:1607.06450, 2016.

[4] Hill F, Cho K, Korhonen A. Learning distributed representations of sentences from unlabelled data[J]. arXiv preprint arXiv:1602.03483, 2016.

[5] Gan Z, Pu Y, Henao R, et al. Learning generic sentence representations using convolutional neural networks[C]//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. 2017: 2390-2400.

[6] Conneau A, Kiela D, Schwenk H, et al. Supervised learning of universal sentence representations from natural language inference data[J]. arXiv preprint arXiv:1705.02364, 2017.

[7] Cer D, Yang Y, Kong S, et al. Universal Sentence Encoder[J]. arXiv preprint arXiv:1803.11175, 2018.

[8] Bowman S R, Angeli G, Potts C, et al. A large annotated corpus for learning natural language inference[J]. arXiv preprint arXiv:1508.05326, 2015.

[9] Li J, Luong M T, Jurafsky D. A hierarchical neural autoencoder for paragraphs and documents[J]. arXiv preprint arXiv:1506.01057, 2015.

[10] Le Q, Mikolov T. Distributed representations of sentences and documents[C]//International Conference on Machine Learning. 2014: 1188-1196.

[11] Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality[C]//Advances in neural information processing systems. 2013: 3111-3119.

[12] Arora S, Liang Y, Ma T. A simple but tough-to-beat baseline for sentence embeddings[C]//ICLR. 2016.

[13] Henderson M, Al-Rfou R, Strope B, et al. Efficient natural language response suggestion for smart reply[J]. arXiv preprint arXiv:1705.00652, 2017.

[14] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//Advances in Neural Information Processing Systems. 2017: 6000-6010.

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

[16] Manning C, Surdeanu M, Bauer J, et al. The Stanford CoreNLP natural language processing toolkit[C]//Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations. 2014: 55-60.

[17] Kingma D P, Ba J. Adam: A method for stochastic optimization[J]. arXiv preprint arXiv:1412.6980, 2014.

[18] Glorot X, Bengio Y. Understanding the difficulty of training deep feedforward neural networks[C]//Proceedings of the thirteenth international conference on artificial intelligence and statistics. 2010: 249-256.