

基于外部记忆单元和语义角色知识的文本复述判别模型*

李天时¹, 李琦¹, 王文辉¹, 常宝宝¹

(1.北京大学, 北京市海淀区 100871)

摘要: 文本复述判别是一个重要的句子级语义理解应用。本文提出了一个轻量级的基于记忆单元的单层循环神经网络模型, 并结合语义角色标注知识帮助进行英文文本复述判别。我们使用单层的循环神经网络模型减缓由于网络层数过多加重的梯度消失和梯度爆炸问题, 易于训练。并且利用外部记忆单元和语义角色知识帮助存储两句话中不同层级的语义联系。我们的模型在英文评测语料 Microsoft Research Paraphrase Corpus 测试集上 F 值为 84.3%。实验表明, 语义角色标注知识确实可以帮助文本复述判别, 并且我们的轻量级模型达到了同多层次神经网络模型相近的效果。

关键词: 文本复述判别; 语义角色; 记忆单元; 循环神经网络

中图分类号: TP391

文献标识码: A

Paraphrase Identification with External Memory and SRL Knowledge

Tianshi Li¹, Qi Li¹, Wenhui Wang¹, Baobao Chang¹

(1.Peking University, Haidian District, Beijing 100871, China)

Abstract: Paraphrase identification is an important sentence semantic understanding application. In this paper, we presented a lightweight memory based recurrent neural network with semantic role features to do this task. Our single layer recurrent network alleviates the gradient disappearance and gradient explosion which aggravate by multilayer neural networks. Our memory and semantic role features can help paraphrase identification. On the test set of Microsoft Research Paraphrase Corpus, we reached 84.3% F1 score. Experiments show that semantic roles are helpful for paraphrase identification and our lightweight model gets competitive result compared with multilayer neural network models.

Key words: Paraphrase Identification; Semantic Role; Memory; Recurrent Neural Network

1 引言

文本复述判别任务旨在判断给定的两句话是否表达相同的含义, 该任务本质上是识别两句话之间的语义相关程度。在文本复述判别任务中: 若两句话是复述关系, 则定义两句话是等价的; 若两句话不是复述关系, 则定义两句话是不等价的。图 1 中给出了等价和不等价的句对实例。上面一组句子是等价的, 下面一组句子是不等价的。第一组句子中两句话在词汇表达上虽然不同 (尤其是加粗部分), 但是整体上两句话表达出了相同的含义, 表达了“美国参议院要发起对伊拉克战前严格的情报准备工作”。第二组句子中虽然两句话的主干表达了相同的含义, 但是由于第一个句子比第二个句子多出了一部分细节描述(句子中加黑部分), 所以二者在语义上不是完全相同, 是不等价的。识别复述关系对于自动问答[13]、抄袭检测[14]、机器翻译评价[15]等不同的任务都有重要的作用。

* **收稿日期:** **定稿日期:**

基金项目: 973 课题; 自然科学基金 (61273318)

作者简介: 李天时 (1992—), 男, 硕士, 自然语言处理; 李琦 (1990—), 男, 硕士, 自然语言处理; 王文辉 (1993—), 男, 硕士, 自然语言处理; 常宝宝 (1971—), 男, 副教授, 自然语言处理。

图 1 英文文本复述判别句子实例

The Senate Select Committee on Intelligence is preparing a blistering report on **prewar intelligence** on Iraq.

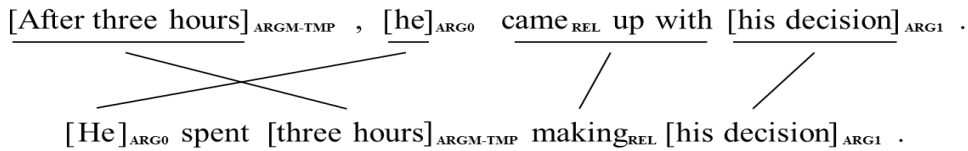
等价: American **intelligence leading up to the war** on Iraq will be criticised by a powerful US Congressional committee due to report soon, officials said today.

SOME %%NUMBER%% jobs are set to go at Cadbury Schweppes , the confectionery and drinks giant , **as part of a sweeping cost reduction programme announced today** .

不等价: Confectionery group Cadbury Schweppes has warned of further cuts to its %%NUMBER%% -strong UK workforce .

目前做法可以分为基于特征的方法和基于多层次神经网络模型的方法, 两类方法都是把该任务当作是二分类问题处理。不同之处在于, 基于特征的方法把重点放在使用不同的特征捕获两句话的语义联系, 基于多层次模型的方法对句子进行建模, 在结合特征之上通过不同层次的模型挖掘更多不同层级的句子语义联系。前人的研究方法有如下的优点: 1) 基于特征的研究方法[1, 4, 8]通过机器翻译评价指标和 N 元组等特征就可以识别出两句话基本的语义联系, 进而取得不错的效果; 2) 基于多层次神经网络模型的研究方法[9, 10, 11, 12]可以

图 2 互为复述关系句子实例及相关语义角色标注。



通过不同层次的网络自动学习到词汇、短语和句子片段等不同层级上的语义联系。而前人的工作中还存在一些不足: 1) 基于特征的研究方法中虽然使用了机器翻译评价指标等特征, 但是这些特征不能提供足够的不同层次语义知识。语义角色标注和文本复述判别同为语义理解任务, 二者有许多联系, 语义角色知识可以帮助识别两句话中的不同层级的语义联系, 而前人研究忽略了这个信息。2) 多层次的神经网络需要使用不同层次上的网络学习不同层级的语义知识, 比如 Socher 等人[9]使用的树结构自动编码器或是 Yin 等人[12]使用的两层以上的卷积神经网络。这导致模型总层数多, 容易加重梯度消失和梯度爆炸的问题, 进而加大训练难度。

图 2 展示了一对互为复述关系的句子。从这两句话中我们可以看出: 1) 两句话在不同层次(词汇、短语、片段)上都有对应语义关联。比如 *came up with* 和 *make* 在表层用词不一样, *came up with* 有想出、提出、追赶上等含义, *make* 更是有开始、做、制造等不同含义, 而在这两个句子中, 结合上下文的语境, 它们表达了提出这个相同的语义。2) 对于 *came up with* 和 *make* 这两个动作的中心动词来说, 两句话中的相关的语义角色也在语义组块和角色类型上有强烈的对应关系。因此, 学习不同层次语义联系, 以及结合语义角色知识进行复述判别是十分有用的。

因此, 本章设计了一种轻量级单层记忆单元神经网络模型, 该模型有如下优点:

- 利用外部单层记忆单元来存储两句话中的不同层级语义联系, 缩减整个模型网络层数, 降低梯度消失和梯度爆炸的问题, 模型易于训练。
- 加入语义角色知识作为特征, 利用对应动词语义角色在两句话间的联系帮助复述判别。

2 文本复述判别模型

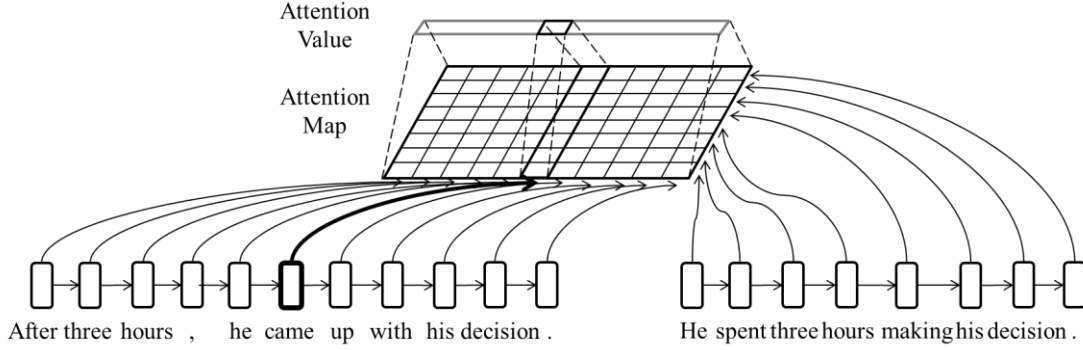
2.1 基于注意力机制的神经网络模型

本节先介绍本章实验使用的基础模型架构, 基于注意力机制(Attention)的单层循环

神经网络复述判别模型。该模型由如下两个模块组成：生成句子向量表示和利用逻辑斯蒂回归模型（LR）进行复述判别。

首先是生成句子向量表示的模块。图 2 中展示了利用注意力机制生成句子向量表示的部分架构。给定两个句子 s_1 和 s_2 ，两个句子的词向量序列分别为 $A = a_1, a_2, \dots, a_{l_1}$ 和 $B = b_1, b_2, \dots, b_{l_2}$ ，其中 l_1 和 l_2 分别是两句话的长度。利用长短记忆单元网络（LSTM）[16] 分别对两句话进行建模，生成新的句子向量表示序列 \bar{A} 和 \bar{B} 。对于一句话中第 t 个词来说，经过

图 3 基于注意力机制（Attention）的循环神经网络模型部分架构。图中展示的注意力得分（Attention Value）是针对左边句子，根据注意力矩阵（Attention Map）按行求和之后需归一化得到的。右边句子的注意力得分可以类似的将注意力矩阵按列求和后归一化得到。



LSTM 层后输出向量表示如下：

$$\bar{a}_t = LSTM(a_t, \bar{a}_{t-1}) \quad (8)$$

$$\bar{b}_t = LSTM(b_t, \bar{b}_{t-1}) \quad (9)$$

其中，两句话的 LSTM 层在权重上是共享的，我们认为这样可以将两句话经过 LSTM 层后的表示映射到同一语义空间，方便后续识别两句话间的语义联系。在 LSTM 模型生成的新句子序列向量表示的基础上，我们利用如下的软对齐的方式生成每句话固定维度的句子向量表示。首先生成 s_1 和 s_2 之间的注意力矩阵（Attention Map） M ， $M \in \mathbb{R}^{l_1 \times l_2}$ 。对于两句话中的任意两个词的原始词向量对 $\langle a_i, b_j \rangle$ ，我们使用下面的公式来计算注意力矩阵 M 中的每个元素 e_{ij} ：

$$e_{ij} = \cos(a_i, b_j) \quad (10)$$

其中 $\cos(\cdot)$ 是计算余弦相似度的函数。该矩阵表明了两句话词语之间的语义相关性。之后，我们通过公式 11 和 12 计算一句话中每个词语对另一句话的重要程度权重：

$$\alpha_i = \sum_{j=1}^{l_2} e_{ij}, \forall i \in [1, \dots, l_1] \quad (11)$$

$$\beta_j = \sum_{i=1}^{l_1} e_{ij}, \forall j \in [1, \dots, l_2] \quad (12)$$

根据一句话中每个词语对另一句话的重要程度权重，我们就可以计算出考虑另一句话相关性的定长句子向量表示：

$$\hat{a} = \sum_{i=1}^{l_1} \frac{\exp(\alpha_i)}{\sum_{k=1}^{l_1} \exp(\alpha_k)} \bar{a}_i \quad (13)$$

$$\hat{b} = \sum_{j=1}^{l_2} \frac{\exp(\beta_j)}{\sum_{k=1}^{l_2} \exp(\beta_k)} \bar{b}_j \quad (14)$$

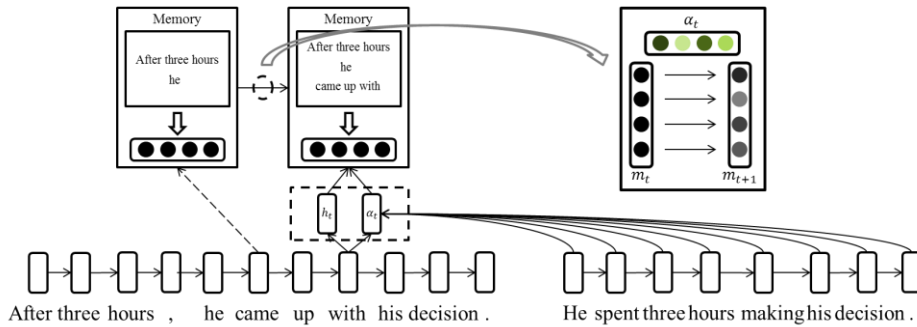
其中 \hat{a} 和 \hat{b} 是通过句子自身经过 LSTM 后单词向量根据另一句话权重加权求和得到。直观上来讲，新生成的句子向量表示是将原句子中同另一句话中相关性强的部分抽取求和得到的，本质上讲是抽取两句话相关性强的部分得到的向量表示。

然后是 LR 模型复述判别模块。在该模块中，我们利用上一模块中生成的句子向量表示拼接上前人使用过的有效特征作为 LR 模型的输入。在具体实现时，我们并不是直接使用两

句话的句子向量表示作为输入特征的一部分，而是使用如下公式计算句子向量之间的余弦相似度 $\cos(\hat{a}, \hat{b})$ 作为输入特征。我们认为，由于两句话的句子向量在同一语义空间，余弦相似度可以更直观的反映两个句子向量的相关程度，进而反映两句话的语义相关程度。不仅如此，我们还使用 LSTM 层输出的句子结尾单词向量表示 \bar{a}_{l_1} 和 \bar{b}_{l_2} 作为原始句子的向量表示并计算二者的相似度。所以最终该模块可以归纳为如下公式：

$$y = LR([\cos(\hat{a}, \hat{b}); \cos(\bar{a}_{l_1}, \bar{b}_{l_2}); x_f]) \quad (15)$$

图 4 基于单层记忆单元（Memory）的循环神经网络模型部分架构。



其中 x_f 是每一维由一个特征组成的特征向量表示， $y \in [0,1]$ 为最终逻辑斯蒂回归模型（LR）的分类输出。

这个模型有如下特点：仅使用了单层 LSTM，并使用单次 Attention 机制一次性捕获两句话中的相关语义联系。然而这个模型还存在着一个明显的问题，Attention 机制的输入端每个词的向量表示是由单层 LSTM 层生成的，这个向量表示虽然在一定程度上可以学习到前后文的语义依赖，但向量中包含的大部分语义信息还是词汇级别的语义信息。这就导致利用 Attention 机制生成的两句话相关联的向量表示也是以词汇级别的语义关联为主，不能真正学习到不同层次的语义联系。

2. 2 基于外部记忆单元的循环神经网络模型

针对上节基础模型还存在的问题，本节引入外部记忆单元加以改进。本节中的模型只对生成句子向量部分表示进行变动，使用单层记忆单元替换基础模型中的注意力机制。我们利用单层的记忆单元存储一句话中对于另一句来说重要的语义信息，然后利用两个句子各自的记忆单元识别句子间的语义联系。在前人多层次的神经网络中：使用卷积网络的方法[10,11,12]由于每一层卷积窗口大小固定，只能学习有限的上下文知识，需要通过不同层次的网络学习不同层级语义知识；使用自动编码器[9]的方法更是需要依赖树状结构学习语义知识，网络本身就是多层次的。我们使用的单层次记忆单元网络基于 LSTM，由于 LSTM 可以捕获前文的信息，所以随着我们记忆单元的更新，记忆单元中会不断加入相对于另一句话重要的语义知识，这些知识又同记忆单元之前存储的语义知识融合，形成不同层次的（词汇、短语、句子片段）语义知识存储单元。

单层记忆单元如图 3 所示。记忆单元是一个独立的向量，每个句子都有一个独立的记忆单元，该记忆单元随着单向循环神经网络在在句子中的前向传播进行更新。以图 3 为例，左边句子传播到 *came* 这个词时，对应左边句子的记忆单元中已经发现句子片段 *After three hours* 和词汇 *he* 同右边句子有强烈的语义联系，因此已经将它们语义联系转化为向量表示。当句子传播到 *with* 时，读入该词后发现同前两个词组成的短语 *came up with* 和右边句子有明显的语义联系，则将这个语义联系也存储到记忆单元向量中。在我们的记忆单元向量中，向量的每一位本质上都表示一个不同的特征，我们可以利用向量 α_t 控制记忆单元中的每一位进行不同的更新操作， α_t 中的每一位就对应着记忆向量中每一位的更新开关。按位进行不同

更新的策略就使得记忆单元向量就可以在每位上存储不同层次上的语义联系。每一次记忆单元更新对应的输入 h_t ，该向量是单层 LSTM 输出的词向量经过空间映射后得到的新向量表示。我们使用 h_t 而不直接使用 LSTM 层输出的向量作为更新记忆单元的输入也是为了使记忆单元中可以存储更丰富的不同层次的语义联系，而不单是词汇级别的联系。

上述记忆单元的工作方式通过下面一系列公式进行自动更新。同上一节一样，给定两句话 s_1 和 s_2 以及他们的词向量序列 $A = a_1, a_2 \dots, a_{l_1}$ 和 $B = b_1, b_2 \dots, b_{l_2}$ ，我们先用 LSTM 对句子建模，建模后输出的向量序列分别为 $\bar{A} = \bar{a}_1, \bar{a}_2 \dots, \bar{a}_{l_1}$ 和 $\bar{B} = \bar{b}_1, \bar{b}_2 \dots, \bar{b}_{l_2}$ 。我们在 LSTM 层之上，加入了外部记忆单元，该记忆单元会随着句子序列在 LSTM 上的正向传播进行不断的更新。以句子 s_1 为例，句子序列到第 t 个词时的记忆单元 m_t 通过如下公式进行更新：

$$m_t = m_{t-1} \odot (1 - \alpha_t) + \alpha_t \odot h_t \quad (16)$$

其中 $m_t \in \mathbb{R}^d$ ， $h_t \in \mathbb{R}^d$ 是基于 LSTM 的输出 \bar{a}_t 得到的向量， $\alpha_t \in \mathbb{R}^d$ 是门向量(Gate Vector)，用来控制保留多少原始记忆单元中信息并读入多少词 t 中信息。 h_t 和 α_t 的计算方法如下：

$$h_t = \tanh(W_m \bar{a}_t + b_m) \quad (17)$$

$$\alpha_t = \max_{1 < j < l_2} score_{j,t}$$

(18)

$$score_{j,t} = \text{sigmoid}(U_m[\bar{a}_t; \bar{b}_j]) \quad (19)$$

其中 W_m ， b_m 和 $U_m \in \mathbb{R}^{d \times 2d}$ 均为网络参数。门向量 α_t 是通过词 t 同 s_2 中的每一个词计算一个相应的权重，然后将生成的权重序列通过最大池化操作生成的。直观上来讲，就是当前词 t 同 s_2 中每个词计算其关联的重要程度，然后取每一维中的最大值最为其最重要程度的参考，然后根据该对 s_2 的重要程度选择写入多少到 s_1 的记忆单元中。句子 s_2 的记忆单元可以使用类似的公式进行计算得到。因此，通过上述公式，我们对每个句子都生成了一个记忆单元，里面存储了该句子相对于另一句话来说不同层级上的重要语义知识，然后再利用这些语义知识进行最终复述判别。

在 LR 模型复述判别模块中，我们使用两个句子的记忆单元，再加上 LSTM 层输出序列的最终向量作为总的句子向量特征表示，希望来利用原有句子表示的同时强化两句话中关联部分的语义联系。同上一节一样，我们不是直接使用句子向量作为 LR 模型的输入，而是先分别计算对应句子向量之间的余弦相似度，然后将余弦相似度作为输入的一部分。

2.3 加入语义角色特征

通过图 2 的例子可以看出语义角色知识也可以帮助文本复述判别挖掘不同层次的语义联系。因此，我们在基于单词记忆单元的循环神经网络模型基础上，通过加入句子中的语义角色知识进一步挖掘更多的不同层次语义关联。

首先，是抽取句子语义角色特征。考虑到一句话中包含多个动词，但是不是所有动词及其语义角色都对句子语义联系识别有帮助作用，尤其是句子中同另一句话语义不相关的动词。我们利用词向量去计算两句话间每个动词对的余弦相似度。预训练的词向量表示包含了词语在文章上下文中的语义信息，使用它计算出余弦相似度越大的动词对语义联系越相近。实验中，我们使用词向量间余弦相似度最大的动词对进行两句话的语义角色标注。我们认为这种做法可以在利用对识别句子语义联系最有用的语义角色特征的同时过滤了语义角色标注知识带来的噪声。具体步骤如下：1) 使用词性标注器识别句子中的动词。2) 找出两句话中词向量余弦相似度最大的动词对。3) 利用该动词对和语义角色标注器对两句话进行语义角色标注。

其次，是利用上述方法抽取语义角色作为复述判别模型输入特征的一部分。对于一句话中的第 t 个词，其原始词向量表示为 $x_t \in \mathbb{R}^{d_{word}}$ ，其语义角色标签对应的特征向量表示为 $f_t \in \mathbb{R}^{d_{srl}}$ 。我们将语义角色特征向量同词向量拼接起来，作为新的单词 t 的向量表示 \tilde{x}_t ，作

为 LSTM 层的输入。

$$\tilde{x}_t = [x_t; f_t] \quad (20)$$

2. 4 Dropout 策略

考虑到复述判别的训练集较小，深层的神经网络容易产生过拟合的问题，我们在 LSTM 层的输出时使用 Dropout 策略来缓解过拟合的问题。Dropout 策略指的是在训练阶段将向量中的每一维按照一定的概率 p 屏蔽，即进行清零操作，然后在测试阶段将该向量的每一维按照 p 的比例进行缩小。原有 LSTM 层按照公式 8 和公式 9 进行计算，加入 Dropout 之后，计算方式如下：

$$\bar{a}_t = Dropout(LSTM(a_t, \bar{a}_{t-1}), p) \quad (21)$$

$$\bar{b}_t = Dropout(LSTM(b_t, \bar{b}_{t-1}), p) \quad (22)$$

其中 p 表示按照 p 大小的概率进行向量按维清零操作。

2. 5 训练策略

本章中使用的最终分类模型是逻辑斯蒂回归模型。给定训练样例：

$$T = (x^{(i)}, y^{(i)}) \quad (23)$$

其中 $x^{(i)}$ 表示第 i 个训练句子对， $y^{(i)} \in \{0,1\}$ 表示第 i 个训练样例正确分类结果。我们的网络模型可以使用如下的公式概括：

$$y_{pred}^{(i)} = EM-LSTM(x^{(i)}, \theta) \quad (24)$$

其中 $y_{pred}^{(i)} \in [0,1]$ 是模型对训练样例 $x^{(i)}$ 的预测结果，EM-LSTM (External-Memory LSTM) 指的是我们加入记忆单元的网络模型， θ 是模型参数集合。

给定训练数据和模型预测结果，我们可以使用极大似然法估计模型参数，其中设模型预测结果是 1 和 0 的概率分别为如下公式：

$$P(Y^{(i)} = 1|x^{(i)}) = y_{pred}^{(i)}, P(Y^{(i)} = 0|x^{(i)}) = 1 - y_{pred}^{(i)} \quad (25)$$

则其似然函数为：

$$\left[y_{pred}^{(i)} \right]^{y^{(i)}} \left[1 - y_{pred}^{(i)} \right]^{1-y^{(i)}} \quad (26)$$

对于整个训练集来讲，总的对数似然函数为：

$$J(\theta) = \sum_i y^{(i)} \log y_{pred}^{(i)} + (1 - y^{(i)}) \log(1 - y_{pred}^{(i)}) \quad (27)$$

我们训练时使用随机梯度下降 (Stochastic Gradient Descent) 算法进行梯度更新。

3 实验

3. 1 实验设置

文本复述判别语料我们使用的是微软提供的 MSRP 语料，该语料中训练集包含 4508 个句子对，测试集包含 1720 个句子对。抽取语义角色特征时训练使用的语料是英文语义角色标注评测语料 CoNLL-2005，使用的是 Wang 等人[17]提出的语义角色标注模型，词性标注器是斯坦福大学的 CoreNLP 工具包¹。针对句子中每个词的语义角色形式也同 Wang 等人一样使用 IOBES 标记形式。的初始词向量我们使用的是 Glove 词向量²。

¹ 斯坦福大学自然语言处理工具：Stanford CoreNLP。下载地址：<https://stanfordnlp.github.io/CoreNLP/>。

² Glove 词向量是由斯坦福大学提供了一种英文词向量表示。下载地址：

本章中提出的复述判别模型参数具体设置如下：词向量维数 100；语义角色特征向量维数 60；LSTM 隐层和输出层向量维数为 100；LSTM 层数为 1；Memory 向量维数 100；Dropout 中 p 值取 0.5；训练学习率为 0.05；实验中采用 mini-batch 训练策略，batch 大小为 20。

我们的实验模型也使用了前人在深度学习模型中使用过的特征，并加入到了逻辑斯蒂回归分类层。我们使用的特征分为两种：第一种是 Yin 等人[12]在他们的卷积网络模型中也使用的八种机器翻译评价指标特征；第二种是 Socher 等人[9]自动编码器模型中使用的识别两句话中数字是否相同的特征。我们使用的特征和上面两种深度学习方法使用的特征在数量级上相同。

3. 2 实验结果和分析

表 1 展示了本章中提出的模型在 MSRP 语料测试集上的效果。其中：LSTM 是只使用一层 LSTM 的基础模型；LSTM+Attention 是本章中提出的基于注意力机制的 LSTM 基础模型；EM-LSTM 是本章提出的加入单层记忆单元的 LSTM 模型；EM-LSTM+SRL 在记忆单元基础上加入语义角色特征的模型。可以看出，随着模型的不断改进，实验效果也不断的提升。LSTM 作为所有实验的对照组，在加入注意力机制后，效果有一定的提升，说明注意力机制确实可以抽取出句子中词汇级别的语义联系。当使用我们提出的单层记忆单元来取代原有注意力机制时，效果进一步提升，说明我们使用的记忆单元确实可以识别并存储两句话中更丰富的不同层次的语义联系。在加入语义角色特征后，实验结果在两种指标上都又有提升，这说明同为语义理解任务的语义角色知识确实可以帮助复述判别。

表 1 本章提出的模型在 MSRP 测试集上的实验结果。

方法	精确率 (%)	F 值 (%)
LSTM	75.6	83.3
LSTM+Attention	76.0	83.4
EM-LSTM	76.7	83.5
EM-LSTM+SRL	77.2	84.3

表 2 展示了在 MSRP 语料测试集上同前人方法比较的结果。其中：Majority 表示按照训练集中占比最大的复述关系（两句话是复述关系）进行分类的结果；RAE 为 Socher 等人[9]使用递归自动编码器模型识别的结果；A-LSTM 是 Yin 等人[12]在实验部分提到的基于注意力机制的 LSTM 模型；ABCNN 是 Yin 等人[12]提出的基于注意力机制的多层次卷积神经网络模型；EM-LSTM 是本章中提出模型的最好效果。同 Socher 等人提出的 RAE 相比，我们的模型达到的效果超过了他们利用自编码器结构上不同层次特征表示进行复述判别的效果。我们的实验结果虽然不如 Yin 等人的多层次卷积模型效果，但是也达到了一个有竞争力的效果。考虑到我们只使用了单层记忆单元存储句子间的语义联系，而他们模型最好的效果使用了两层卷积网络，并且每个卷积层可以拆分为纵向的三个子网络层，因此我们网络层数是远少于他们的。Yin 等人在实验部分也给出了只使用一层 CNN 的效果，精确率为 78.1%，F 值为 84.1%。可以看出，在 F 值指标上还不如我们模型的效果，这说明我们单层记忆单元的模型确实是有效的，可以在一定程度上达到同多层次模型一样的效果，学习到不同层级的语义联系。

表 2 在 MSRP 测试集上的实验结果同前人比较。

方法	精确率 (%)	F 值 (%)
Majority	66.5	79.9

RAE	76.8	83.6
A-LSTM	77.1	84.0
ABCNN	78.9	84.8
EM-LSTM+SRL	77.2	84.3

4 前人工作

基于特征的文本复述判别方法通常利用不同的特征来获得不同的语义联系,进而进行复述判别。Wan 等人[1]结合机器翻译评价指标 BLEU 值和一些基于依存关系和树上编辑距离的特征,使用支持向量机(Support Vector Machine)进行分类。Mihalcea 等人[2]结合点互信息(Pointwise Mutual Information)、浅层语义分析和基于 WordNet 测量的词语相似度来生成一个统一的判断是否有复述关系的指标。Qiu 等人[3]建立了一个可以检测两句话之间的不同点的架构,利用不同点的显著程度来判断两句话是否有复述关系。Kozareva 等人[4]利用最长公共子序列(Longest Common Subsequence)、N 元组和 WordNet 等特征,尝试了支持向量机、k 近邻、最大熵等分类器。Fernando 等人[5]基于两句话的所有词建立了一个词语的相似度矩阵。Das 等人[6]建立了一个两句话之间的对齐模型,并结合逻辑斯蒂回归分类器(Logistic Regression Classifier)进行分类。Blacoe 等人[7]利用神经网络语言模型生成词语向量表示,然后通过求和方式生成句子的向量表示特征,进而利用句子向量计算相似度。在诸多基于特征的方法中,效果最好的是 Madnani 等人[8]在 2012 年提出的,利用多个不同的机器翻译指标作为特征的分类模型。

虽然基于特征的方法可以捕获一定不同层次的语义信息,但是效果好坏依赖于特征的设计,并且捕获的语义层次有限。随着深度学习的发展,许多研究人员在使用特征的方法的基础上,将原有训练模型使用神经网络代替,并不断加深神经网络层数旨在自动学习到更多的不同层次语义联系。2012 年,Socher 等人[9]使用自动编码器分别对两句话构建贪心的树结构网络,然后计算两个句子网络中不同层次节点之间的余弦相似度,最后将生成的矩阵通过池化的操作得到固定长宽的矩阵作为逻辑斯蒂回归的输入进行分类。更多的研究使用卷积神经网络来处理该任务,通常是两个句子分别建立多层卷积层,然后使用最后一层的结果作为 LR 的输入进行分类。2014 年,Hu 等人[10]利用卷积神经网络对句子进行建模。之后,Yin 等人[11]针对不同层级的卷积层,引入两句话的相似度交互信息。更进一步,Yin 等人[12]在之前卷积神经网络工作的基础上,加入注意力机制,利用两个句子相同层次上的网络向量计算注意力矩阵,通过该矩阵辅助指导不同层卷积层之间的更新。

在前人的工作中,基于特征的方法主要利用机器翻译等指标作为特征,而不同的机器翻译评价指标同质化严重,没有蕴含足够多的语义知识。基于多层次神经网络模型的方法为了挖掘不同层次的语义联系,使用了不同层次的网络,在训练时加重了梯度消失和梯度爆炸的问题。前人的卷积神经网络模型单是卷积层就使用两层以上,这增加了模型训练的难度。

5 总结

针对英文文本复述任务,本文提出了一种结合语义角色知识的轻量级单层记忆单元网络模型。该模型通过语义角色特征捕获了更多语义联系,并使用单层记忆单元存储了不同层次的语义联系,起到了压缩多层次网络模型层数的作用,便于模型训练。通过实验,我们的轻量级模型达到了与多层次神经网络模型接近的效果。

参考文献

- [1] Stephen Wan, Mark Dras, Robert Dale et al. "Using dependency-based features to take the "parafarce" out of paraphrase". In: Proceedings of the Australasian Language Technology Workshop, 2006.

- [2] Rada Mihalcea, Courtney Corley, Carlo Strapparava et al. "Corpus-based and knowledge-based measures of text semantic similarity". In: AAAI, 2006: 775-780.
- [3] Long Qiu, Min-Yen Kan and Tat-Seng Chua. "Paraphrase recognition via dissimilarity significance classification". In: Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, 2006: 18-26.
- [4] Zornitsa Kozareva and Andrés Montoyo. "Paraphrase identification on the basis of supervised machine learning techniques". In: Advances in natural language processing. Springer, 2006: 524-533.
- [5] Samuel Fernando and Mark Stevenson. "A semantic similarity approach to paraphrase detection". In: Proceedings of the 11th Annual Research Colloquium of the UK Special Interest Group for Computational Linguistics, 2008: 45-52.
- [6] Dipanjan Das and Noah A Smith. "Paraphrase identification as probabilistic quasi-synchronous recognition". In: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1, 2009: 468-476.
- [7] William Blacoe and Mirella Lapata. "A comparison of vector-based representations for semantic composition". In: Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, 2012: 546-556.
- [8] Nitin Madnani, Joel Tetreault and Martin Chodorow. "Re-examining machine translation metrics for paraphrase identification". In: Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2012: 182-190.
- [9] Richard Socher, Eric H Huang, Jeffrey Pennington et al. "Dynamic Pooling and Unfolding Recursive Autoencoders for Paraphrase Detection." In: NIPS, 2011: 801-809.
- [10] Baotian Hu, Zhengdong Lu, Hang Li et al. "Convolutional neural network architectures for matching natural language sentences". In: Advances in neural information processing systems, 2014: 2042-2050.
- [11] Wenpeng Yin and Hinrich Schütze. "Convolutional Neural Network for Paraphrase Identification." In: HLT-NAACL, 2015: 901-911.
- [12] Wenpeng Yin, Hinrich Schütze, Bing Xiang et al. "Abcnn: Attention-based convolutional neural network for modeling sentence pairs". arXiv preprint arXiv:1512.05193, 2015.
- [13] Erwin Marsi and Emiel Krahmer. "Explorations in sentence fusion". In: Proceedings of the European Workshop on Natural Language Generation, 2005: 109-117.
- [14] Paul Clough, Robert Gaizauskas, Scott SL Piao et al. "Meter: Measuring text reuse". In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, 2002: 152-159.
- [15] Chris Callison-Burch. "Syntactic constraints on paraphrases extracted from parallel corpora". In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2008: 196-205.
- [16] Hochreiter Sepp and Jürgen Schmidhuber. "Long short-term memory." Neural computation, 1997: 1735-1780.
- [17] ZhenWang, Tingsong Jiang, Baobao Chang et al. "Chinese Semantic Role Labeling with Bidirectional Recurrent Neural Networks." In: EMNLP, 2015: 1626-1631.

作者联系方式：李天时 北京市丰台区东高地红星北里甲 1 栋 6 单元 301 100076
13717631219 lts_417@hotmail.com