软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software, [doi: 10.13328/j.cnki.jos.000000]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

基于 Bi-LSTM 和 Attention 的文本关系分类网络*

姚昕智[1,2]

[1](华中农业大学 信息学院 湖北 武汉 226019)

[2](湖北省生物信息重点实验室(华中农业大学),湖北 武汉 226019)

通讯作者: 姚昕智, E-mail: xinzhi_bioinfo@163.com

摘 要: 文本分类一直是自然语言处理（NLP）领域中的重要语义处理任务，文本分类的准确性对下游任务，例如关系抽提，命名实体注释都具有重要意义。在传统的文本关系分类方法中，通常基于两类传统思路，一类是基于其他人工处理方法，例如依存树解析和语法注释的方法来获取文本的高维度特征，另一类方法更倾向于利用深度人工神经网络自动对文本的特征进行抽提。在对文本特征进行自动抽提的过程中，特征抽提网络往往需要解决两方面的问题，一方面是句子前后上下文的远距离信息难以被捕获并且学习，另一方面是句子对分类结构具有重要影响的单词存在于句子中的不同位置。所以改模型在利用 Bi-LSTM 来自动捕获单词前后上下文信息，同时利用词嵌入级别的 Attention 机制来获取句子中分散的重要语义信息。模型分别在长短文本分类任务上做了测试，均取得了很好地效果。

关键词: 文本关系分类，Bi-LSTM，Attention，特征抽提

中图法分类号: TP311

Text relation classification network based on Bi-LSTM and Attention

CXINZHI Yao[1,2]

[1](College of Information, Huazhong Agricultural University, Wuhan 226019, China)

[2](College of Information, Huazhong Agricultural University, Wuhan 226019, China)

Abstract: Text classification has always been an important semantic processing task in the field of natural language processing (NLP). The accuracy of text classification is important for downstream tasks such as relationship extraction and named entity annotation. In traditional text relationship classification methods, there are usually two types of traditional ideas. One is based on other manual processing methods, such as dependency tree parsing and grammatical annotation methods to obtain high-dimensional features of text. Deep artificial neural networks automatically extract the features of the text. In the process of automatically extracting text features, feature extraction networks often need to solve two problems, one is that the long-distance information of the context before and after the sentence is difficult to capture and learn, and the other is that the sentence is important for the classification structure. The affected words exist in different places in the sentence. Therefore, the modified model uses Bi-LSTM to automatically capture the context information before and after the word, while using the Attention mechanism of the word embedding level to obtain the important semantic information scattered in the sentence. The model has been tested on long and short text classification tasks, and both have achieved good results.

Key words: Text relation classification, Bi-LSTM, Attention, feature extraction

文本关系分类是查找文本中所包含的名词对之间的语义关系的任务，该任务在 NLP 各子领域中都十分重

* 基金项目: 国家自然科学基金(00000000, 00000000);华中农业大学湖北省武汉市生物信息重点实验室开放课题 KFKT00000000)

Foundation item: National Natural Science Foundation of China (00000000, 00000000); State Key Laboratory for Novel Software Technology (Nanjing University)开放课题 (KFKT00000000)

收稿时间: 0000-00-00; 修改时间: 0000-00-00; 采用时间: 0000-00-00; jos 在线出版时间: 0000-00-00

CNKI 在线出版时间: 0000-00-00

要，例如在舆情分析领域的文本情感分类，在问答系统设计领域问题的分类准确性对下游系统至关重要。同时文本关系分类对许多 NLP 的任务也十分有用，例如信息提取，问题解答。

传统的关系分类方法利用人工注释文本组成单词中所包含的高级特征协助关系分类，通常是基于规则的模匹配，并且能达到很好的性能。这些方法的缺点之一是许多传统的 NLP 工具被用于抽提文本中的高级特征，例如利用 Stanford core NLP 工具对文本的句法依存关系进行分析，或是通过 NCBI 提出的工具 PubTator 来对 PubMed 的医学文献文本进行命名实体识别用于协助下游文本关系分类，但这样的工具抽提文本高级特征的方法，一方面导致计算成本的增加，同时增加了其他船舶错误，另一个缺点是手动设计功能并加入到后续文本分类任务重十分耗时，并且由于不同训练数据集的覆盖率较低，因此在通用化方面表现不佳。

最近，深度学习方法的崛起让很多人尝试自利用深度网路自动抽提文本特征，该方法提供了减少手工特征数量的有效方法。但是，很多基于深度学习的方法依然依赖于词法资源，例如 WordNet 或是其他 NLP 系统（如依赖解析器和 NER）来获得高级功能。

该模型利用一种新的网络结构用于关系分类的神经网络。这一模型利用双向长短时记忆网络结合注意力机制来自动捕获文本中分散的重要语义信息。改模型没有利用词汇资源或者依赖于其他 NLP 系统产生的高级特征，它可以使用带有注意力机制的双向长短时记忆网络自动捕获句子中最重要的语义信息。

改模型分别在长文本和短文本的文本关系分类任务中都取得了很好地效果，分别在验证集达到了 89%和 90%的准确性，89.6%和 90.0%的 F1 分数。

1 相关工作

在最近这些年以来，已经有很多用于文本关系分类的方法被提出，他们大多是都基于模式匹配，并且应用额外的 NLP 系统来解析文本所具有的高级词汇特征并用于后续分类任务。在 Rink 和 Harassbagin 年 2010 提出的文本关系分类系统中，利用了从外部语料库衍生的许多功能来利用支持向量机分类器。最近，深度人工神经网络常常被用于自动学习文本的基本特征，并且已经在改模型中使用。最有代表性的开创性工作是有 Zeng 等人在 2014 年完成的，他们利用卷积神经网络进行关系分类。在后续的其他工作中证明 CNN 不适合学习远程语义信息，所以有人尝试利用基于递归神经网络（RNN）进行远距离特征的抽提。Zhang 和 Wang2015 年提出一项工作，该工作用双向 RNN 从原始文本数据中学习关系模式，尽管双向 RNN 可以同时学习前后的上下文信息，但是由于梯度消失和梯度爆炸的问题上下文但范围收到了限制。为了克服这个问题，Hochreiter 和 Schmidhuber（1997）引入了长短时记忆单元（LSTM）。另一个相关的工作是 Yan 等人提出的 SDP-LSTM 模型，改模型利用文本中两个实体之间的最短依存路径（SDP），然后使用 LSTM 网络学习 SDP 的信息，同时将原始文本视为一个序列，将其与 SDP 进行拼接同时用于后续的文本关系分类任务。

2 模型

在本结中，我会介绍该模型以及模型中所包含的五个部分.

(1) 输入层，用于输入改文本关系分类模型的文本。

(2) 嵌入层：将每个单词从高维度的文本表示映射到低维向量。

(3) LSTM 层：利用双向长短时记忆网络从步骤二计算自动抽提文本的高级特征。

(4) Attention 层：利用注意力机制，对（3）中产生的每个单词计算权重，得到一个权重向量，并通过将权重向量相乘，将词嵌入通过双向长短时记忆网络计算得到的每个单词级别的特征合并成为句子级别的特征向量。

(5) 输出层：句子级别的特征向量，最终通过 SoftMax 韩式用于文本关系分类。

2.1 词嵌入层

在利用深度学习方法除以文本数据的时候，与处理图像数据不同，为了使高维的的文本表示可以进行深度网络中的计算，我们将高维度的文本映射到低纬度的空间，这样的方法在自然语言处理领域称为词嵌入方法（Word Embedding）。在将高维度数据映射的同时，单词低纬度的向量表示往往能够携带更多的语义信息，例如在 Word2Vec 和 Glove 论文中所提及的方法，这样的语义信息将有利于我们的后续任务进一步抽提句子级别的文本信息。

2.2 Bi-LSTM 层

LSTM 最初被提出用于解决循环神经网络 RNN 在学习文本上下文远距离关系时由于反向传播造成的梯度爆炸及梯度小时问题，所以在 RNN 的每一个单元中，单独增加了门的设置，在能够更好地传递文本远距离上下文信息的同时，还能够通过细胞状态进一步保留，删除，输出特定信息，并且通过实验证明，长短时记忆网络在处理文本上下文远距离信息时能够有更好的表现。

LSTM 的公式推到如下：i_t 和 g_t 称为输入门，决定当前 LSTM 单元储存什么新的信息，例如当前所处理句子的新的主语。f_t 称为遗忘门，来决定当前时间状态的 LSTM 单元需要忘记什么信息，例如当处理到一句新的句子时，需要对之前处理的句子的主语进行遗忘。c_t 用于更新当前时间 LSTM 单元的状态，而 h_t 用于决定什么信息通过当前时间状态输出并且传递给下一时间状态。

$$
\begin{aligned}
i_t &= \sigma(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i) \\
f_t &= \sigma(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f) \\
g_t &= \tanh(W_{xc}x_t + W_{hc}h_{t-1} + W_{cc}c_{t-1} + b_c) \\
c_t &= i_t g_t + f_t c_{t-1} \\
o_t &= \sigma(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_t + b_o) \\
h_t &= o_t \tanh(c_t)
\end{aligned}
$$

而当我们在利用 LSTM 网络学习句子上下文远距离信息的时候，由于单向的 LSTM 网络往往只关注于特定单词某一方向的上下文，而忽略了对另一放下上下文的特征抽提，所以在该模型中，我们利用双向的长短时极易网络来同时对一个单词的前后上下文进行特征抽提，从而更好地学习上下文信息，并且我们将不同方向的长短时记忆网络每个时间状态的输出进行拼接作为某一特定时间点的单词编码，如下列公式。

$$
h_i = [\overrightarrow{h_i} \oplus \overleftarrow{h_i}]
$$

2.3 Attention 层

注意力机制在 2016 年被提出，基本思想是在人为处理图片或者文本数据的时候会有不同的侧重点，而往往我们只需要对关键侧重点进行提取分析就可以很好地得出我们所想要的结果，同时还能有效地减少误差。在我们的网络中，我们使用单词级别的 Attention 机制，旨在自动分布在句子中不同位置的关键语义，即给特定单词分配更高的注意力权重，从而获得更好句子级别编码，用于后续任务， Attention 机制的公式如下。

公式中 H 为双向 LSTM 对每个单词重新编码所构成的矩阵，表示为 $H=[h_1, h_2, ..., h_T]$，其中 T 为句子长度，α 为利用一个全连接层来逼近注意力机制中计算相似度的过程，r 则表示通过将注意力权重矩阵和句子编码相乘获得句子级别的新编码。

$$M = \tanh(H)$$
$$\alpha = softmax(w^T M)$$
$$r = H\alpha^T$$

2.4 output 层

在改模型的输出层，我们将注意力机制所获得的句子级别编码作为 softmax 分类网络的输入，并且将结果中概率最大的标签作为网络的预测结果。同时在该模型中，我们使用传统的分类神经网络损失函数：负对数函数作为损失函数用于训练神经网络。

$$J(\theta) = -\frac{1}{m}\sum_{i=1}^{m} t_i \log(y_i) + \lambda \|\theta\|_F^2$$

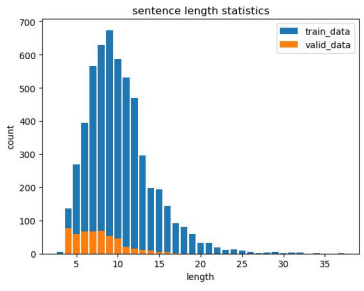
$$\hat{y} = \arg\max_y \hat{p}(y|S)$$

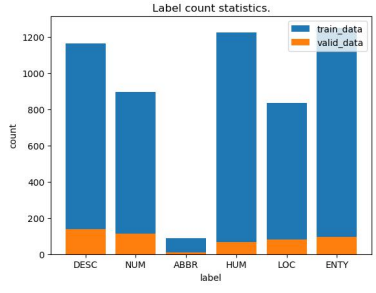
3 实验

3.1 数据

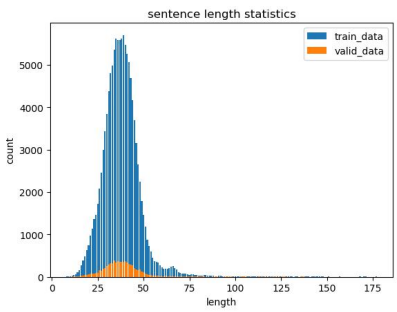
在我们的实验中，为了分别验证该模型在短文本数据和长文本数据中的文本分类能力，我门分别从网络上收集了两套数据集。

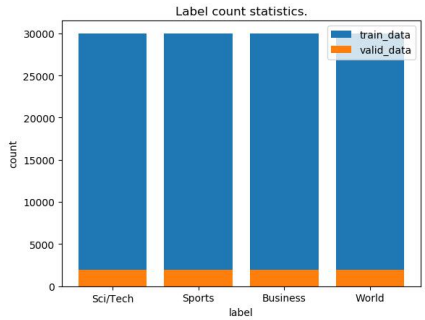
其中短文本分类数据集是关于问题分类的六分类问题，其中包含 NUM, LOC, HUM, DESC, ABBR, ENTY 六个标签，训练集包含 5,450 个句子，验证集包含 500 个句子。为了更好的利用数据集，我们分别对数据集的长度分布以及标签平衡情况进行了统计，可见自短文本数据集中，文本长度主要分布在 5 到 20 之间，而标签的分布存在不平衡问题，这可能对我门后续的网络训练造成影响，如下图。





同样的，我们从（http://groups.di.unipi.it/~gulli/AG_corpus_of_news_articles.html）收集了 Antonio Gullli's corpus 数据集，这是一个关于新闻主题分类的四分类任务数据集。长度分布主要在 25 到 70 之间，而 4 个标签的数量相同，更利于后面的网络训练，如下图，其中橙色表示验证集而蓝色表示训练集。

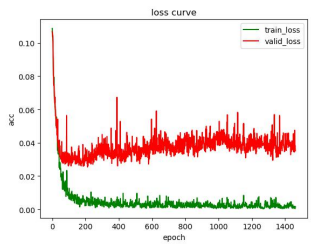


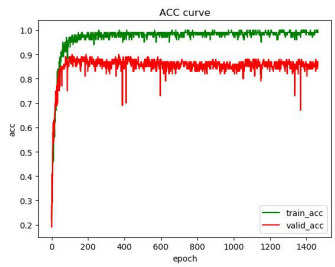


3.2 实验结果

我们分别在短文本数据集和长文本数据集上做了训练和验证。

在短文本数据集中，训练集的损失函数能够很平稳的下降并收敛，但是验证集的损失会存在波动，同样的精准率在验证集上的效果也同在训练集上的数据差狠毒，我们推断这是由于训练集和验证集的标签不平衡造成的，如下图，其中绿色表示训练集而红色表示验证集。在短文本数据集上我们最后训练集上准确率能达到 98%，而验证集上准确率能达到 89%。整个模型在验证集上能达到 0.911 的 precision score，0.881 Recall Score 和 0 .896 的 F1-score，如下表。

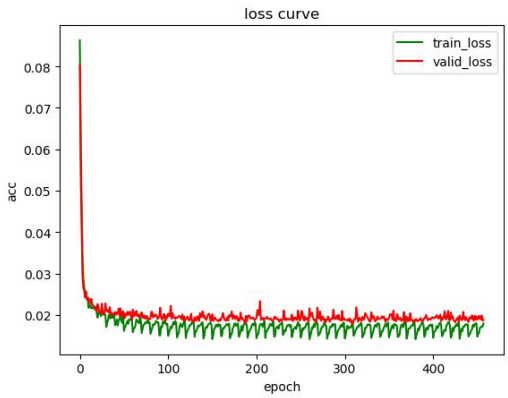


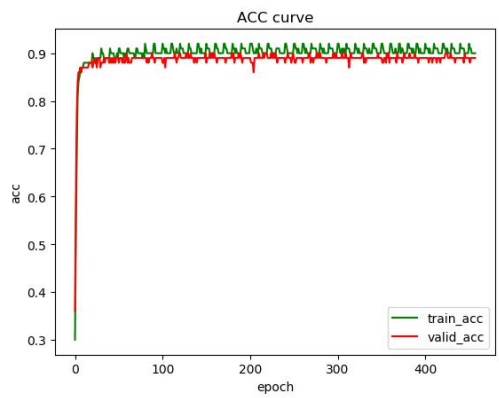


短文本	Accurcy
训练集	98%
验证集	89%

短文本	Precision	Recall	F1
BL-A	0.911	0.881	0.896

同样的在长文本数据集中，模型在训练集和验证集上的损失都能够很好地下降并且收敛，同时训练集和验证机上的准确度基本相同，可见数据集的标签平衡对模型的训练非常重要，如下图，红色表示训练集而绿色表示验证集。而在训练集上的准确率为 91%，验证机上的准确率为 89%，整个模型在验证集上的 Precision score





长文本	Accurcy
训练集	91%
验证集	89%

长文本	Precision	Recall	F1
BL-A	0.906	0.903	0.903

4 总结

该 Bi-LSTM 和 Attention 结合的模型在长短文本数据集上都能够获得很好地效果，然而越来越多的新模型被提出自动抽提文本高级特征且用于下游任务。技术无止境，希望之后能利用更合理的深度学习模型来应用于自然语言处理任务，从而达到更好的效果。

本项目所有代码均原创，存于：https://github.com/YaoXinZhi/Bi-LSTM-Attention

References:

[1] Zhou, Peng, et al. "Attention-based bidirectional long short-term memory networks for relation classification." Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). 2016.

[2] Zhu Xiaoying, Lai Shaohui, Lu Keda. Application of LSTM in News Classification [J]. Journal of Wuzhou University, 28 (06): 16-26.

[3] Fan Lu, Fuyuan Hu, Junyu Shen, et al. Multi-label classification method for images with double LSTM structure [J]. Journal of Suzhou University of Science and Technology (Natural Science Edition), 2018 (3): 79-84.