基于 LSTM 的中文语义分类研究

尹红梅[1]，陈琛[2]
1.扬州海关技术服务中心，江苏 扬州 225800
2.江苏旅游职业学院，江苏 扬州 225800

摘　　要： 经典的语义分类算法，在中文上直接使用，会出现不同的问题。将英文语句切分成单词，根据单词出现的次序，英文语境中，影响不大。中文的字与字之间，前后是关联的。一个字出现在不同位置，含义是不同的。中文语义，根据区位码与汉字建立一一对应的关系。实验显示，基于分句为字，区位编码这两个原则，本文提出的中文语义分析算法取得了良好效果。

关 键 词： 深度学习；语义匹配；固定编码；神经网络

Research on Chinese Semantic Classification Based on LSTM

Yin Hongmei[1], Chen Chen[2]
1.Yangzhou Customs Technical Service Center, Yangzhou, Jiangsu 225800
2.Jiangsu Vocational College of Tourism, Yangzhou, Jiangsu 225800

Abstract: Classic semantic classification algorithms, when used directly in Chinese, may encounter different problems. Splitting English sentences into words, according to the order in which they appear, has little impact in the English context. Chinese characters are related before and after each other. A word appearing in different positions has different meanings. Chinese semantics establish a one-to-one correspondence between location codes and Chinese characters. The experiment shows that the Chinese semantic analysis algorithm proposed in this paper has achieved good results based on the principles of dividing sentences into words and encoding location.

Keywords: deep learning; semantic matching; fixed coding; neural network

引言

近年来，语义分类成为研究热点，1995年提出的长短程记忆方法取得良好效果。前向网络预测下一个单词，使用回归网络可以根据所有单词实现预测。英语语境中，LSTM分成两个步骤：首先，将训练集中单词按出现次序编码。虽然测试集中也做同样操作，同样的单词由于出现次序不同，编码也不同。第二步，为了保证语句长度相同，长语句尾部截除，短语句尾部增加空格补齐语句长度。语义是连续的。文章是把相互关联的词连接而成的。LSTM使用长程协方差学习和分类。它从数据时间戳中学习到长程依赖。文本输入LSTM后，先转换成数据序列，将单词编码。为得到更好结果，网络中增加单词嵌入层。它将词典中的单词，变成一个数值向量，而不采用标量索引。这种嵌入，采集单词语义细节。通过一个向量算法将单词关系模型化，语义相似的单词就有了相似向量。

本文在传统LSTM基础上，根据中文语境，做了改进。构建了新的LSTM模型，遵循“分句为字，固定编码”原则，对文本语句实现内容分析，取得较好实验结果。

一、相关工作

论文综述了语义场景分类研究，分析了现有分类算法的性能，并指出了语义分类中存在的问题，特别关注室内场景分类的挑战[1]。针对目前Deep Web分类研究中所采用的Post-query查寻探测方法缺乏语义支持的问题，提出一个基于本体的语义查询探测分类方法训练文本经过DeBERTa模型神经网络后，得到原始文本的特征向量表示，再与解释序列的特征向量进行融合，以实现极短文本的层次分类[2]。提出了一种结合多元语义特征和图卷积神经网络(GCN)的短文本分类模型，将语义特征同短文本一起构建一个多元异构图，利用GCN学习短文本更深层特征，进而实现短文本分类[3]。提出了一种文本分类模型SEB-GCN，其在文本词共现图的基础上加入了句法文本图与语义文本图，提高模型的分类效果[4]。设计一种面向分类网络的视觉语义解释模型，对飞机类别进行分类[5]。构建基于MASK机制的词类别预测任务对预训练模型BERT进行微调，以学习单词与类别的关系，方法在公开数据集上，分类准确率得到提高[6]。提出了解耦的共享语义空间嵌入方法，改进了利用标签语义信息的方法，利用预训练模型中的先验知识增强标签层次结构信息. 实验结果表明，该方法在公开数据集上优于目前最先进的多标签文本分类模型[7]。针对文本语义特征，语境语义特征和标记实体语义特征，建立多重语义融合机制，，实现关系分类模型[8]。该模型提升了关系分类模型的性能。[9]自提出了一种基于标签语义注意力的多标签文本分类方法，依赖于文档的文本和对应的标签，使用双向长短时记忆获取每个单词的隐表示，能

够有效地捕获重要的单词，并且其性能优于当前先进的多标签文本分类算法。[10]提出一种利用层级标签语义信息引导的极限多标签文本分类模型提升策略，在训练和预测过程中给予模型层级标签引导的弱监督语义指导信息，能够有效提升现有模型性能。[11]提出了一种创新点语义识别与分类方法，将科技文摘按照句法和语义功能进行6分类算法处理，然后对6分类算法结果进行了类与句子位置的数量分布统计分析。实验结果表明，这种方法算法简便，分类精度高，普适性好。[12]提出一种结合深度卷积神经网络和集成分类器链的多标记图像语义标注方法，利用深度卷积神经网络学习图像的高层视觉特征，基于获取的视觉特征与图像的语义标记集训练集成分类器链，与一些当前国际先进水平的方法相比，文中方法的鲁棒性更强，标注结果更精确。[13]提出基于语义依存分析的图网络文本分类模型，对文本进行语义依存分析编码，快速挖掘语义依存信息，使得网络更快地收敛。

二、算法改进

算法分成三个阶段：阶段1：收集数据，获取汉字。阶段2：数据预处理和特征抽取。阶段3：数据分割，模型训练和验证。

卷积可用如下公式表示：

$$f_l^k(p,q)=\sum_c\sum_{x,y} i_c(x,y)\cdot e_l^k(u,v)$$

其中，输入矩阵是一个卷积核。借助LSTM基本结构，我们自建了一个模型。输入为二分输入，隐含层宽度为10，全连接层宽度为4。与英文处理相比[14]，中文算法改进了语句分割和数据编码。在GB2312-80汉字库中，每个汉字都有固定编码，分为区码和位码。将语句分割成单个汉字，并用区位码对汉字编码。

三、实验结果分析

我们收集有价值的数据总共有6000条。涉及酒店预订、客房服务、投诉建议和其他评价。为了验证算法，我们设计了三个维度的实验。分别从输入宽度、层数和数据量的变化，检验算法的有效性，最后，根据实验，选定最佳参数，得到优化模型。

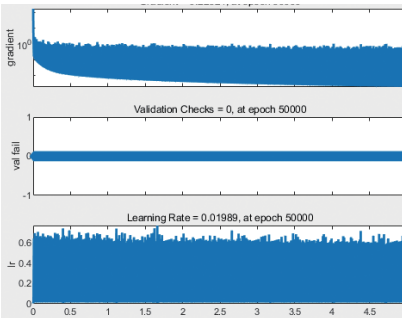
初始情况，隐含层数设定为100，输入宽度设定为20，数据量取100。首先，将输入神经元数目增加到200，观察实验结果。

经过196轮训练，学习率达到0.0073052，网络快速收敛。根据观察，神经元数目增加，对训练不是总是有益的。当输入神经元数目增加到一定程度时，梯度下降速度不再明显，错误率反而增加了。训练到70轮时，学习率下降36%，训练时间减少142%。实验结果与理论分析基本吻合。

隐含层数目增加，层间交互增加，网络复杂程度增加。因为一层相当于一个函数，增加层数计算更加精确，计算复杂度提高，耗费的时间更多[15]。训练到458轮时，梯度衰减逐渐平缓。平均学习率下降，显示神经网络层数目对网络模型有很大影响。

模型训练需要大数据支持。我们将6000个数据分成三部分，70%用于训练，20%用于验证，10%用于测试。也就是用4200条语句用于训练，1200条语句用于验证，600条语句用于测试。

实验显示，验证正确率达到95.17%，测试准确率达到100%。遵循“断句为字，固定编码”原则，LSTM网络能够成为中文语义分类的理想算法。



> 图1 训练结果展示图

四、结束语

从语句中分割出短语，是一件很难的工作，而且，不同的规则和语境，对短语分割有很大影响。即使采用人工分割，由于不能设计出统一规则，分割结果千差万别。汉字作为计算机ASCII基本集上的编码集，每个汉字都是由内码和形码编码生成的。汉字集是固定的二维矩阵，能够实现汉字与数值的一一对应。每个汉字有固定的释义，用汉字作为语义基元，能够减轻语句切割的难度，又能保障不同语句的中的汉字编码统一。基于深度学习的语义分析算法，用于中文语境，是否合适，是否正确，有待于进一步深入研究。本文提出的解决方案，在中文语义理解上是一种有益的尝试，从实验结果看，是可行的。未来，在中文语境进行语义理解和分析，可以探索更多的深度学习算法。

参考文献

[1]顾广华，韩晰瑛，陈春霞，等. 图像场景语义分类研究进展综述[J]. 系统工程与电子技术，2016,38(4):936-948.

[2]吕强，宋玲，马军，等. 基于本体的Deep Web语义分类研究[J]. 山东建筑大学学报，2010,25(2):118-124.

[3]鲁富宇，冷泳林，崔洪霞. 基于多元语义特征和图卷积神经网络的短文本分类模型[J]. 河南科学，2024,42(5):625-630.

[4]孙红，陆欣荣，徐广辉，等. 融合语义和句法依存分析的图卷积新闻文本分类[J]. 中文信息学报，2023,37(7):91-101.

[5]吕学强，赵兴强，贾智彬，等. 面向分类网络的视觉语义解释模型[J]. 计算机工程，2023,49(11):220-230.

[6]贾晨晓，欧阳丹彤. 多重语义融合的关系分类模型[J]. 吉林大学学报（信息科学版），2023,41(1):50-56.

[7]肖琳，陈博理，黄鑫，等. 基于标签语义注意力的多标签文本分类[J]. 软件学报，2020,31(4):1079-1089.

[8]王嫄，徐涛，王世龙，等. 层级标签语义引导的极限多标签文本分类策略[J]. 中文信息学报，2021,35(10):110-118.

[9]温浩. 科技文摘创新点语义识别与分类方法研究[J]. 情报学报，2019,38(3):249-256.

[10]李志欣，郑永哲，张灿龙，等. 结合深度特征与多标记分类的图像语义标注[J]. 计算机辅助设计与图形学学报，2018,30(2):318-326.

[11]范国凤，刘璟，姚绍文，等. 基于语义依存分析的图网络文本分类模型[J]. 计算机应用研究，2020,37(12):3594-3598.

[12]陈肖磊，曾银龙，韦波，等. LBS的数据语义分类研究[J]. 地理空间信息，2009,7(2):36-38.

[13]陈昊飚，张雷. 融合语义解释和DeBERTa的极短文本层次分类[J]. 计算机科学，2024,51(5):250-257.

[14]林呈宇，王雷，薛聪. 标签语义增强的弱监督文本分类模型[J]. 计算机应用，2023,43(2):335-342.

[15]孙坤，秦博文，桑基韬，等. 基于共享语义空间的多标签文本分类[J]. 计算机工程与应用，2023,59(12):100-105.