计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2023,32(12):74−83 [doi: 10.15888/j.cnki.csa.009320]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

改进 DPCNN 分类模型在金融领域长文本的应用①

王　婷, 梁佳莹, 杨　川, 何松泽, 向　东, 马洪江

(成都信息工程大学 计算机学院, 成都 610225)
通信作者: 马洪江, E-mail: mhj68@163.com

摘　要: 为了解决金融领域文本分类算法稀缺, 以及现有算法无法充分提取文本中词与词的关系、长距离依赖关系和深层次特征信息的问题, 提出了一种改进卷积自注意力模型的文本深度关系抽取算法. 该算法在改进的深度金字塔卷积神经网络 (DPCNN) 中引入自注意力, 并联合双向门控神经网络 (BiGRU) 模块建立文本分类模型, 解决了针对金融领域长文本的长距离依赖特征信息和词与词之间关系特征信息的提取问题, 实现文本中深层次特征信息和上下文语义信息联合抽取功能. 在 THUCNews 短文本与长文本数据集上分别进行实验, 实验结果表明, 所提方法与 BERT 等方法相比, 在评价指标上有显著提高. 在自制金融长文本数据集上的对比实验表明, 与其他模型相比, 该算法模型的准确率和 *F*1 值更高. 通过一系列实验可以证明, 该算法模型能够更准确地完成针对金融长文本的分类任务.

关键词: 自然语言处理; 长文本分类; 金融文本; 特征联合提取; 自注意力机制

引用格式: 王婷,梁佳莹,杨川,何松泽,向东,马洪江.改进 DPCNN 分类模型在金融领域长文本的应用.计算机系统应用,2023,32(12):74–83. http://www.c-s-a.org.cn/1003-3254/9320.html

Improved DPCNN Classification Model for Long Texts in Finance

WANG Ting, LIANG Jia-Ying, YANG Chuan, HE Song-Ze, XIANG Dong, MA Hong-Jiang

(School of Computer Science, Chengdu University of Information Technology, Chengdu 610225, China)

Abstract: To solve the scarcity of text classification algorithms in finance and the inability of existing algorithms to adequately extract word-to-word relations, long-distance dependency, and deep feature information in texts, this study proposes a text depth relationship extraction algorithm based on improved convolutional self-attention model. The algorithm introduces self-attention in a modified deep pyramidal convolutional neural network (DPCNN) and builds a text classification model jointly with bi-directional gated neural network (BiGRU) module to solve the problem of extracting long-distance dependency feature information and word-to-word relationship feature information for long texts in finance. Then the joint extraction function of deep feature information and contextual semantic information in texts is realized. Experiments on THUCNews short text and long text datasets show that the proposed method has significant improvement in evaluation indexes compared with BERT and other methods. The comparison experiments on the dataset of homemade financial long texts show that the accuracy and *F*1 value of the algorithm model are higher compared with other models. A series of experiments demonstrate that the algorithmic model can perform the classification task against financial long texts more accurately.

Key words: natural language processing (NLP); long text classification; financial text; joint feature extraction; self-attention mechanism

① 基金项目: 四川省科技厅重点研发项目 (2022YFG0375, 2023YFG0099, 2023YFG0261, 23ZDYF0473, 23ZDYF0181); 南充生物医药产业技术研究院项目 (22YYJCYJ0086); 四川省科技服务业示范项目 (2021GFW130)
收稿时间: 2023-05-24; 修改时间: 2023-06-28; 采用时间: 2023-07-07; csa 在线出版时间: 2023-10-19
CNKI 网络首发时间: 2023-10-20

文本分类是自然语言处理 (natural language processing, NLP) 最重要的任务之一, 旨在将不同类别的无序文本内容按类别分类. 早期的文本分类大多依靠人工来完成, 此方法耗时耗力. 但随着科技的快速发展和计算机的广泛应用, 文本信息已全面电子化, 海量文本扑面而来. 面对庞大的文本数据, 传统的人工分类方法已无法在合理的时间范围内处理并提炼以帮助使用者决策. 基于此, 研究者们利用计算机来进行文本分类任务, 提出各种文本分类算法, 不仅加快了分类速度, 还提高了分类的准确率.

随着机器学习的发展, 涌现出一系列用特征工程和浅层分类模型解决大型文本特征提取和分类问题的方法, 主要包括传统的机器学习方法和深度学习方法. 传统机器学习方法有朴素贝叶斯算法[1]、K 邻近算法[2]和支持向量机[3], 该类方法依赖人为制作的特征进行训练, 存在纬度高、数据稀疏、模型简单、特征表达能力弱等缺点, 导致分类模型过拟合、特征信息提取不全面等问题. 深度学习中的神经网络模型主要有卷积神经网络 (convolutional neural networks, CNN)[4]、循环神经网络 (recurrent neural networks, RNN)[5]、长短时记忆网络 (long short term memory, LSTM)[6]和门控循环单元 (gated recurrent unit, GRU)[7], 该类模型在特征提取方面能得到很好的效果. 余本功等人[8]将 CNN 应用于中文短文本的分类, 上层的数据有助于下层的分类, 分类效果远好于传统的基于机器学习的分类方法. 与 CNN 相比, RNN 能处理顺序信息, 在长句的分类中效果更好, 但是存在梯度损失和长期依赖的问题. Zhou 等人[9]提出结合双向 LSTM 和二维最大池的文本分类算法, 使用 LSTM 模型提取上下文特征, 规避了 RNN 存在的问题. 然而, LSTM 模型训练带来两个明显的缺点: 梯度爆炸和复杂的模型结构引起训练时间过长. Zulqarnain 等人[10]使用词嵌入和 GRU 对文本进行分类, 通过词嵌入技术将文本中的词转化为向量, 再利用 GRU 提取到词与词之间的上下文语义. 与传统的 RNN 和 LSTM 相比, GRU 有效地学习文本语境下的词汇用法, 在分类准确率和错误率方面也取得了较好的效果.

上述机器学习和深度学习算法目前在各领域的文本分类任务中得到广泛应用. 例如: 金宁等人[11]在农业问答问句分类中, 提出基于双向门控循环网络和多尺度卷积神经网络的分类模型 (bi-directional gated recurrent unit_multi-scale convolutional neural network, BiGRU_MulCNN), 该模型由文本预处理层、BiGRU 层和 MulCNN 层组成, 主要利用 BiGRU 获取输入词向量的上下文语义信息, 并构建一个并行的多尺度 CNN 来提取多尺度特征. 从其实验结果来看, 上下文语义信息对文本分类算法有至关重要的影响. 有研究人员将能够提取长期依赖关系的学习技术和改进的自注意力机制应用于文本分类任务, 例如张博等人[12]提出了基于迁移学习和集成学习的医学短文本分类模型, 利用 4 种不同的神经网络分类模型 (CNN、DPCNN、LSTM 和 Self-Attention) 完成医学短文本分类. 其中, 深度金字塔卷积神经网络 (deep pyramid convolutional neural networks, DPCNN) 是一种复杂度较低的词级别深层 CNN 模型, 相比于 CNN, 其中增加了一个卷积模块和一个负采样层, 将整个模型的计算量限制在不到两倍卷积块的范围. 然而, 单一的模型以及注意力的分散, 使分类效率降低, 在实验部分结合自注意力机制后, 医学文本分类系统的性能提高. Dai 等人[13]提出一个位置自注意力层来生成不同的掩码自注意力, 以及一个位置融合层, 融合位置信息和掩码自注意力, 便于生成含有不同位置信息的句子嵌入, 自注意力机制的引入有效地平衡了不同掩码之间的影响. 新闻文本分类可用于舆情分析等领域, 现存算法较多且有多个公开数据集, 例如: 张海丰等人[14]提出的结合 BERT 特征投影网络 (FPnet) 的新闻主题文本分类方法, 并在不同的新闻数据集上进行了实验, 结果表明, 该方法可以在一定程度上提高分类效果. 而范昊等人[15]从融合上下文特征的角度结合 BERT 词嵌入的新闻标题分类模型, 该模型结合了从双向长短时记忆网络 (bidirectional long short term memory, BiLSTM) 和 TEXTCNN 中提取的上下文特征信息. 从其实验结果看, 该模型可以相对准确地对新闻标题进行分类, 而且错误分类极少. 胥佳仙等人[16]通过使用融合了藏语音节和文件之间的构成关系的图卷积网络构造分类模型, 解决藏文新闻文本的分类问题. 该模型的实验准确率比 Word2Vec+LSTM 高 15.65%, 这填补了文本分类在藏文中的空白.

目前, 文本分类在舆情分析、情感分析和垃圾邮件处理等方面都有应用且各领域文本分类研究正在不断发展, 涌现出诸多模型用于解决文本分类的各种问题. 但在中文文本分类领域中, 依然存在一些挑战. 首先, 对于中文文本数据短语和词级建模问题. 相较于英

文, 中文文本数据更加注重于词汇信息的表达而非单个字符. 例如“闻”和“新闻”, 前者是一个动词, 而后者是一个名词, 所表示的含义也完全不同. 其次, 中文文本数据具有一词多义问题, 且容易忽视输入序列各个位置的字或词对分类结果的贡献程度. 以上两个挑战是中文文本分类将为常见且棘手的问题, 同时这些挑战也存在于金融文本分类领域中.

高效的金融文本分类模型可以应用到消费者评价、金融风险监控和金融股票推荐等典型场景中, 帮助快速分析金融文本, 获取到有价值的信息. 但是, 就目前的研究现状而言, 对于金融文本分类领域, 还存在一些该领域独有的挑战尚未解决. 第一, 对于金融领域的文本, 标注文本数据非常少, 缺乏高质量标注好的数据集; 第二, 金融领域分类体系类别粒度细, 且金融文本具有文本长度较长、文本间较为相似、特征不突出以及文本上下文关联紧密等特点. 现有的通用模型无法直接通过金融文本信息进行准确的分类. 针对上述问题和挑战, 本文构建了一个高质量的中文金融领域长文本数据集, 且提出一种用于金融长文本分类任务的优化卷积自注意力模型 (RoBERTa-WWM-DPCNN-Self-Attention-BiGRU, RDAG). 该模型一方面使用 RoBERTa-WWM 对输入文本进行编码, 以更好地获取短语和词级向量. 另一方面通过对 DPCNN 进行改进并引入自注意力机制用于解决一词多义问题、金融数据集文本长度较长以及金融领域文本数据上下文紧密的特点.

本文主要贡献如下.

(1) 改进深度金字塔卷积神经网络 (DPCNN) 并引入自注意力, 提取金融文本的长距离依赖关系特征, 丰富分类信息.

(2) 融合双向门控神经网络 (BiGRU) 模块, 获取到金融类长文本的深层次特征信息和语义信息, 结合其他特征信息, 对分类结果有积极影响.

(3) 自制金融长文本数据集, 解决金融领域高质量数据集缺失的问题.

1 相关工作

1.1 基于预训练模型进行文本分类

注意力机制通过模仿人类的注意力观察机制, 能够为重要信息分配更多的权重, 达到提取有效信息的作用. 该思想于 20 世纪 90 年代提出, 最早用于视觉图像领域, 2014 年 Mnih 等人[17]将注意力机制与 RNN 结合应用于图像分类, 取得了不错的效果. 同年 Bahdanau 等人[18]将注意力机制应用在机器翻译任务上, 将翻译和对齐同时进行. 同时, 这也是首次将注意力机制应用在 NLP 领域, 随后注意力机制在 NLP 相关问题的算法设计上被广泛应用. 由于注意力机制需要严重依赖外部信息, 无法捕捉数据内部相关性. 基于此, 演变出了自注意力机制. 在文本分类任务中, 自注意力机制主要是通过计算单词间的相互影响关系, 来解决长距离依赖问题. 杨兴锐等人[19]提出结合自注意力的文本分类模型, 利用自注意力机制赋予卷积运算后文本信息的权重, 来提取重要特征信息. 2017 年 Vaswani 等人[20]提出了 Transformer, 其大量使用自注意力机制来学习文本表示. 基于 Transformer 的编码器结构, Devlin 等人[21]提出变换器的双向编码器表示 (bidirectional encoder representation from Transformers, BERT) 预训练模型, 相较于 RNN 更有效, 可以捕捉更长距离的依赖关系, 应用于 NLP 文本分类任务上, 可以极大程度提高分类效果. Liu 等人[22]提出一种鲁棒优化的 BERT 预训练方法 (robustly optimized BERT pretraining approach, RoBERTa), 该方法通过对 BERT 预训练模型进行超参数调优、调大训练批次以及增加新的预训练数据集等操作进行优化, 使其在文本分类任务的分类效果高于原始 BERT 模型. Cui 等人[23]针对 BERT 以字为单位建模, 损失词义的问题, 将全词掩码 (whole word masking, WWM) 应用于中文且与 RoBERTa 结合, 构造出 RoBERTa-WWM (robustly optimised BERT pretraining approach whole word mask, RoBERTa-WWM) 模型, 该模型对于中文 NLP 任务能够更好地编码. 王仁超等人[24]提出 RoBERTa-WWM-TEXTCNN 模型, 将 RoBERTa-WWM 捕获的句子级别的语义信息与 TEXTCNN 捕获的局部信息相结合, 从而获取更丰富的语义信息, 该模型在水电工程施工安全隐患文本分类任务中取得了较好的效果. 上述研究都是通过预训练模型或改进的预训练模型对文本进行编码, 从而更好地提取文档句子级编码以提升分类效果.

1.2 基于 DPCNN 模型进行文本分类

DPCNN 模型[25]是由腾讯人工智能实验室于 2017 年提出的基于词级别 (word-level) 的网络, 通过不断叠加网络抽取长距离依赖关系, 以解决 TEXTCNN 无法通过卷积获得文本长距离依赖关系的问题. 模型

具体结构如图 1 所示. 首先文本通过词嵌入层得到对应的词嵌入矩阵, 该层由 3 个卷积核大小不同的卷积神经网络组成, 其次词嵌入矩阵经过两层等长卷积, 以丰富词嵌入矩阵的语义信息, 最后通过一个 1/2 池化的残差块提高词位的语义. 模型通过引入残差网络是为了解决在训练过程中可能出现的梯度消失和梯度爆炸问题. 同时通过实验证明, 不计成本的前提下, 不断增加网络深度就可以获得更高的准确率. 基于此, 自 DPCNN 论文发布后, 此算法模型得到研究者们的广泛应用. 加米拉·吾守尔等人[26]提出一种基于多卷积核 DPCNN 的维吾尔语文本分类联合模型, 即随机初始化多个相同大小、参数不同的卷积核去提取向量化表示文本的语义依赖, 并与 BiLSTM 和 CNN 串行方式学习到的语义信息进行加强模型对全局和局部信息的理解. Yu 等人[27]建立了用于短文本分类的新型深度 CNN 模型, 打破了传统 CNN 无法提取到有用的文本信息和有效的长距离依赖关系的瓶颈, 实验结果表明, 该模型将词级深度卷积网络应用于短文本分类任务效果优异. 由此可见, 将 DPCNN 引入文本分类任务可以有效提升文本分类效果.

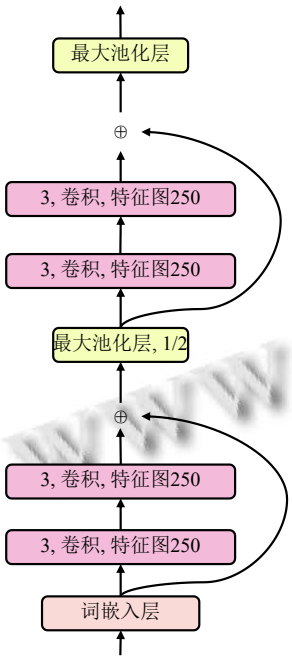

图 1 DPCNN 模型结构图

1.3 基于 BiGRU 模型进行文本分类

在 NLP 任务中, 自然语言序列在结构上存在较为严谨的前后时序, 虽然 GRU 内部的空间虽然可以高效地处理序列时序, 但受限于 GRU 自身单向性的空间结构, 只能捕捉到上文时序语义, 导致全局语义信息的缺失. 因此提出双向 GRU 模型, 即 BiGRU 模型. BiGRU 由两个单向的、方向相反的、输出由这两个 GRU 共同决定的 GRU 组成的神经网络模型. BiGRU 每一时刻的输出都是由这两个单向 GRU 共同决定, 而输入也会同时提供给两个方向相反的 GRU, 解决了 GRU 不能提取到下文时序语义的问题. Liu 等人[28]将 BiGRU 和 CNN 相结合, 提出一种基于字符的文本分类算法, 实现对文本全局和局部语义提取, 但是缺乏文本的局部特征. 基于此, 陈可嘉等人[29]提出一种融合改进自注意力机制的 BiGRU 和多通道 CNN 的文本分类模型, BiGRU 模型提取文本的上下文语义信息和深层次特征信息, 自注意力机制对 BiGRU 层的输出向量进行权重分配, 有效提高文本分类性能.

2 本文模型

本文针对金融领域长文本数据, 提出了一种基于 DPCNN 改进的卷积自注意力文本分类算法 (DPCNN-Self-Attention, DSACNN), 以解决文本长度变化、长距离依赖和注意力分散的问题. 通过实验验证, 本文模型相比于传统深度学习分类模型性能更优, 同时在金融领域长文本分类表现出更优结果; 通过引入自注意力机制, 有效提取词与词之间的关系信息; 融合 BiGRU 模型, 解决文本深层次特征提取不足的问题.

RDAG 模型结构如图 2 所示, 算法具体流程如下: 首先对所输入的文本进行文本预处理, 去除标点符号等无意义字符, 以减少噪声对结果的影响; 然后通过 RoBERTa-WWM 模型对输入文本数据进行编码并获得训练后的特征矩阵; 其次将特征矩阵分别输入到 DSACNN 模型和 BiGRU (bi-directional gated recurrent unit, BiGRU) 模型中分别提取文本长距离依赖信息和文本深层次语义信息; 最后将两个模型得到的特征向量进行拼接, 拼接后的文本特征向量传入分类器中进行分类.

2.1 编码层

金融数据不仅文本数据长而且包含大量专业词汇, 需要模型通过上下文信息进行深度理解. 编码层作为模型的第一部分, 其目的是将输入文本编码为固定长度的向量表示. 相对于传统的词向量编码方法, 例如 Word2Vec、GloVe 在面对长文本复杂的上下文语义关联、不同语境以及一词多义等问题时无法准确表示文

本正确含义. 而添加自注意力机制的预训练语言模型能够解决一词多义问题且能够更丰富表达文本内容, 例如 BERT、RoBERTa 等预训练语言模型. RoBERTa 预训练语言模型通过对谷歌提出的 BERT 预训练语言模型增加 Transformers 的层数, 同时进行了更加精细的调优工作, 使其在处理更长且更多样性的文本数据时更有优势. 但 RoBERTa 采用与 BERT 一样的训练方式, 即 MLM (masked language model) 训练方式, 该方式通过掩盖单个字符从而预测这几个字符. 对于中文来说, 这种训练方式只能得到局部语义信息, 缺乏对短语或词级特征建模. 针对上述情况, 为了更充分地表示文本上下文内容并获取词级语义信息, 本文使用 RoBERTa-WWM 预训练语言模型作为模型编码器. 通过修改 MLM 训练方式, 使用 mask 代替一个完整的词而非字符, 获取词级语义表示, 从而弥补 RoBERTa 预训练语言模型不能获取词级特征信息.

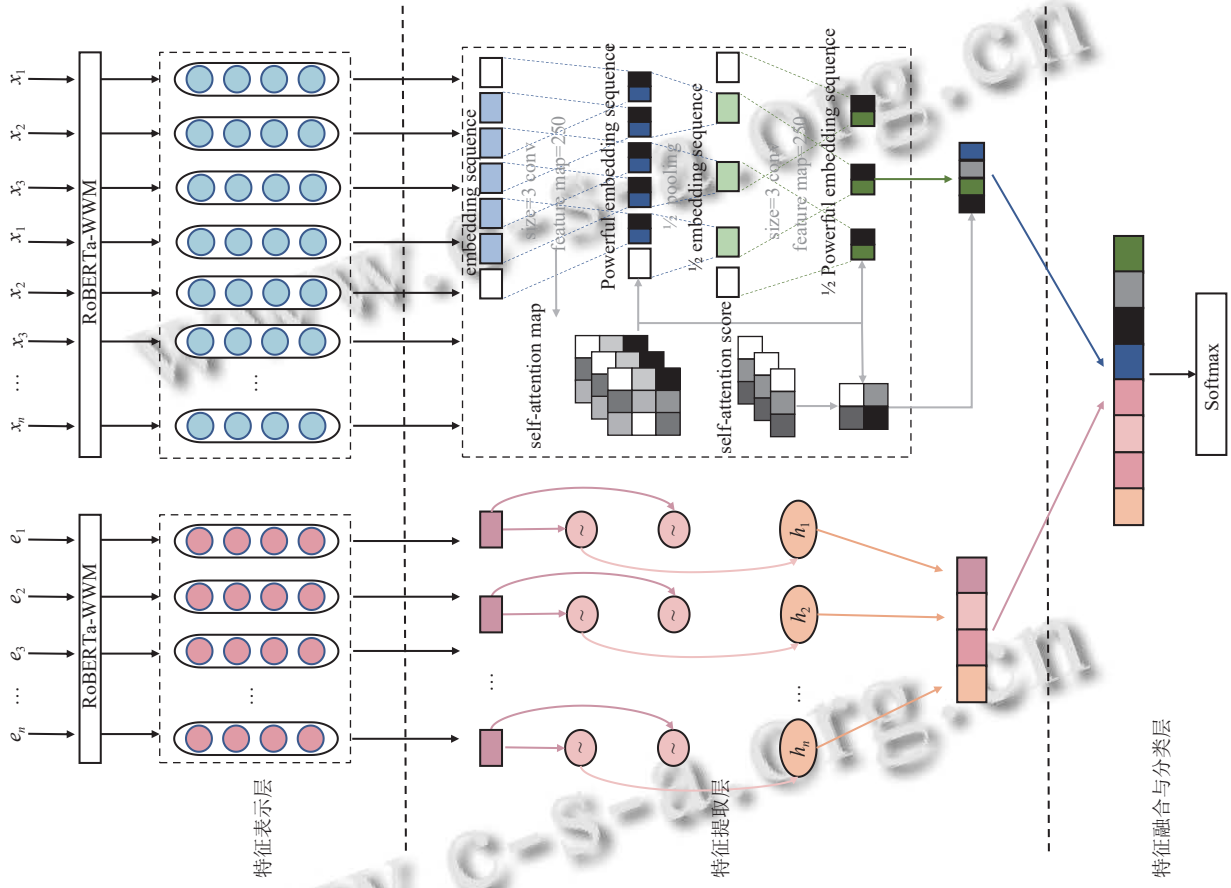

图 2 RDAG 模型结构图

模型具体结构如图 3 所示, 模型嵌入层由 *PE* (position embeddings)、*SE* (segment embeddings) 以及 *TE* (token embeddings) 组成.

$$E_i = PE + SE + TE \tag{1}$$

通过嵌入层得到每个字符的嵌入表示, 随后传入 RoBERTa-WWM 预训练语言模型进行编码.

$$X = (X_1, X_2, \cdots, X_n) = \text{RoBERTa-WWM}\ (E_1, E_2, \cdots, E_n) \tag{2}$$

2.2 DSACNN 模型

为金融领域文本作为一种特殊的文本序列数据, 在文本分类任务中需要模型提取文本的长距离依赖特征. DPCNN 采用一种“深度金字塔结构”的架构, 通过多个重复单元逐渐缩小特征图的宽度, 能够在不损失上下文信息的情况下处理长距离依赖性, 对包含少量词语的短文本更有效. 但由于本文数据集为金融长文本, 模型存在较深的网络结构和大量卷积操作, 可能引发训练时间长和过拟合等问题. 因此本文结合金融长文本的特征对 DPCNN 模型中等长卷积层的数量进行调整, 即减少两层卷积以保留更多的特征信息. 为了更适配该长文本数据集, 利用自注意力机制来建模词语

之间的关系, 从而更好地捕捉文本的语义信息, 填补 DPCNN 模型特征提取不足的缺陷. DPCNN 原始等长卷积层, 部分替换为自注意力机制, 结合自注意力分数通过加权平均的方式综合考虑所有词语的信息. 具体公式见式 (3) 和式 (4):

$$x^t = Regionembedding\left(x^{(t-1)}\right) \tag{3}$$

$$x^c = \mathrm{cov}\left(x^t\right) = W\sigma\left(x^t\right) + b \tag{4}$$

其中, $x^{(t-1)}$是编码层输出的特征矩阵, x^t代表对文本片段进行一组卷积操作后生成的词向量矩阵, W为卷积核的权重参数, σ代表 Sigmoid 激活函数, b代表偏置参数, x^c代表经过一层等长卷积后的输出矩阵. 模型具体结构如图 4 所示.

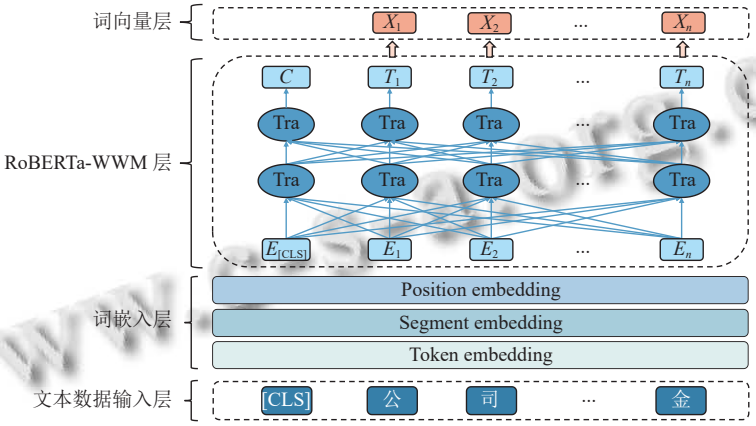

图 3 RoBERTa-WWM 模型结构图

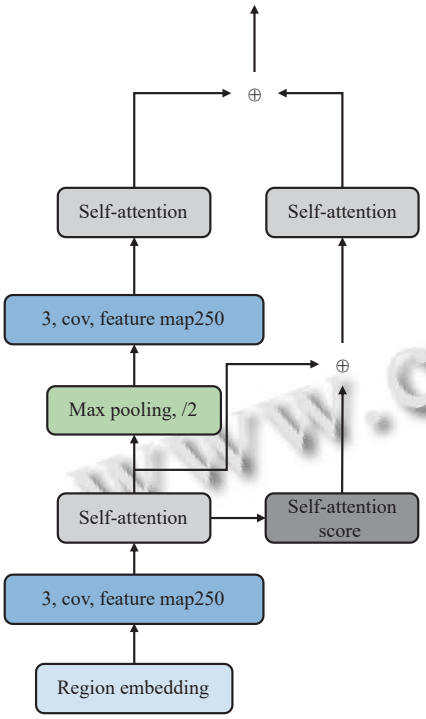

图 4 DSACNN 模型结构图

自注意力机制将词向量分别乘以不同的变换矩阵 W, 得到查询矩阵 (query, Q)、关键字矩阵 (key, K) 和值矩阵 (value, V). Q 与 K 做内积结果为注意力分数 $Score$ 矩阵, 表示 Q 和 K 的相似度.

$$Q = xW_{\mathrm{query}} \tag{5}$$

$$K = xW_{\mathrm{key}} \tag{6}$$

$$V = xW_{\mathrm{value}} \tag{7}$$

$$A = Attention(Q,K,V) = Softmax\left(\frac{\left(QK^{\mathrm{T}}\right)}{\sqrt{d_k}}\right)V \tag{8}$$

$$Score = QK^{\mathrm{T}} \tag{9}$$

其中, W_{query}、W_{key}和W_{value}分别是 3 个可训练的参数矩阵, d_k是线性变换后的维度, A是自注意力机制的输出矩阵. 随后将自注意力机制引入 DPCNN 模型, 将自注意力分数和融合等长卷积的自注意力机制结合, 得到更多样的语义信息.

$$S = Score \oplus A \tag{10}$$

$$x^m = maxpool(A) \tag{11}$$

其中, x^m是大小为 3、步长为 2 的池化层进行最大池化的结果, $\oplus$是拼接函数, S 则是将注意力分数和自注意力机制输出拼接后的输出矩阵.

2.3 BiGRU 模型

BiGRU 模型融合了一个具有前向和后向传播的两

阶段 GRU 模型, 通过双向传播的结构解决上下文影响的问题, 同时考虑当前时间步之前和之后的上下文信息, 更好地提取到序列文本的深层次语义特征信息. 单个 GRU 通过重置门和更新门来控制信息的传递.

其中, 重置门根据上一时刻的隐藏状态 , 来决定过去的信息中有多少需要进行遗忘操作; 更新门则是根据当前时刻 和上一时刻的隐藏单元 , 来决定上一时刻以及当前时刻总共有多少有用信息需要向下传递. 所以 BiGRU 在某个时刻的隐藏状态是通过前向隐藏状态与反向隐藏状态加权求和所得.

$$\overrightarrow{ht} = GRU(Xt, \overrightarrow{ht-1}) \tag{12}$$

$$\overleftarrow{ht} = GRU(Xt, \overleftarrow{ht-1}) \tag{13}$$

$$ht = W t\overrightarrow{ht} + V t\overleftarrow{ht} + bt \tag{14}$$

其中, $GRU(\cdot)$函数将词向量编码为对应的GRU隐藏状态, Wt、Vt分别代表当前时刻双向GRU中前向隐藏状态$\overrightarrow{ht}$和反向隐藏状态$\overleftarrow{ht}$所对应的权重参数, bt则代表当前时刻隐藏状态所对应的偏置参数.

2.4 损失函数

得到带有文本特征矩阵后, 模型需要从特征信息中学习推断出类别和文本特征信息之间的关联. 该模型采用 Rdrop (regularized dropout), 即每个数据样本重复经过带有 Dropout 的同一个模型, 再使用 KL (Kullback-Leibler) 散度 (用来衡量两个概率分布相似性的一个度量指标) 来约束两次的输出尽可能的恒定, 但由于 Dropout 的随机性, 可以近似认为两次的模型略微不同. 具体计算公式见式 (15)–式 (18):

$$L_i = -\log P_\theta(y_i|x_i) \tag{15}$$

$$L_1 = -\log P_\theta^{(1)}(y_i|x_i) - \log P_\theta^{(2)}(y_i|x_i) \tag{16}$$

$$L_2 = \frac{1}{2}[KL(P_\theta^{(1)}(y_i|x_i)|P_\theta^{(2)}(y_i|x_i)) + KL(P_\theta^{(2)}(y_i|x_i)|P_\theta^{(1)}(y_i|x_i))] \tag{17}$$

$$L = L_1 + \partial L_2 \tag{18}$$

其中, $P_\theta(y_i|x_i)$是用于计算文本分类模型的文本数据, $(x_i|y_i)$指的是训练数据, 而$P_\theta^{(1)}(y_i|x_i)$和$P_\theta^{(2)}(y_i|x_i)$分别代表两次进入模型后得到的输出, L_1和L_2则是两部分损失函数的结果, ∂指权值, 最后两部分的加权和L为最终 Loss 值.

3 实验

3.1 实验数据的获取与预处理

为验证模型的通用性和在金融领域长文本的有效性, 实验分别在公开的 THUCNews (THU Chinese text classification) 数据集和自制的金融文本数据集上进行.

THUCNews 公开数据集共有 10 个类别: 0-金融、1-房地产、2-股票、3-教育、4-科技、5-社会、6-政治、7-运动、8-游戏、9-娱乐. 为了验证本文模型对文本长度的特殊性, 分为短文本数据集 (THUCNews-S) 和长文本数据集 (THUCNews-L) 分别进行实验. THUCNews-S 一共 20 万条数据, 由于句子中的符号会给训练带来噪声, 所以在预处理过程中清洗掉标点符号, 处理后各类别句子的平均长度为 16. THUCNews-L 一共 6 万条数据, 预处理过程后各类别句子的平均长度为 760.

自制金融数据集来源于中国研究数据服务平台 (Chinese research data services platform, CNRDS) 中的上市公司年报管理层讨论与分析数据库. 来源的权威性、数据的可靠性及其高质量确保了该模型的实验效果更具有说服力. 在下载好数据包后, 通过“?”“.”“!”等符号对文本进行分句, 对分句后的文本数据进行预处理, 以去除停顿词和标点符号, 从而提高其质量. 经过预处理, 将每篇 30–150 字的文本按以下内容分为 5 类: 0: 非前瞻性语句; 1: 行业发展前景与市场竞争格局; 2: 上市公司未来经营计划; 3: 上市公司资金需求与资金来源; 4: 上市公司发展面临的风险、机遇与对策. 自制金融数据集共计 9421 条数据, 各数据集如表 1 所示.

表 1 数据集信息表

名称	类别数	训练集	测试集	验证集	平均长度
THUCNews-S	10	180000	10000	10000	16
THUCNews-L	10	50000	10000	10000	760
自制金融长文本	5	6281	1570	1570	69

3.2 评价指标

使用准确率 (Accuracy, Acc)、召回率 (Recall, R)、精确率 (Precision, P) 和 $F1$ 值 ($F1$-Score, $F1$) 作为评估指标, 以测试该模型对金融文本数据和中文文本分类的通用性. 具体公式见式 (19)–式 (22):

$$Acc = \frac{TP+TN}{TP+FP+TN+FN} \tag{19}$$

$$R = \frac{TP}{TP+FP} \tag{20}$$

$$P = \frac{TP}{TP+FN} \tag{21}$$

$$F1=\frac{2\times P\times R}{P+R} \tag{22}$$

其中, TP和FP分别代表本身为正样本, 预测后分别为正样本的样本数和预测为负样本的样本数; 而TN和FN分别代表本身为负样本, 预测后分别为负样本的样本数和预测为正样本的样本数.

3.3 实验参数与对比模型

该模型基于 PyTorch 框架, 用于训练的 GPU 是 RTX A5000. 经过多次实验对比, 确定学习率等参数, 参数值设置如表 2 所示.

表 2 参数设置表

参数	数值
词向量维度	768
BiGRU维度	(768, 2)
学习率	0.2
Dropout率	0.1
Epoch	20

为了验证本文模型在中文文本分类任务的有效性以及在自制金融数据做文本分类的优越性. 我们分别在 THUCNews 数据集以及自制金融领域数据集做了对比和消融实验. 本文共选择了 5 种模型与 RDAG 模型进行对比, 分别是 BERT、FastText、TextRCNN、DPCNN 和 EasyDL, 具体介绍如下.

(1) BERT 模型[17]: 由谷歌团队研发出的双向编码器, 只需要一个额外的输出就可以对预训练的 BERT 模型进行微调, 应用到文本分类任务中.

(2) FastText 模型[30]: 由 Facebook 人工智能研究团队基于 Word2Vec 进行改进, 融合多种粒度文本信息、简化网络结构的文本编码方法.

(3) TextRCNN 模型[31]: 融合了 RNN 与 CNN 的优点, 最大化捕捉上下文信息, 同时避免了 RNN 对重点信息位置的限制以及 CNN 固定窗方法中窗口大小对模型的影响.

(4) DPCNN 模型[25]: 由腾讯 AI 实验室提出的可以通过卷积网络与金字塔结构, 充分获取到文本的长距离依赖信息.

(5) 百度 EasyDL 模型: 百度推出的一款操作简单的应用, 用户上传带有标签的文本数据, 直接训练使用.

(6) RoBERTa-WWM 模型[22]: 由 Facebook 团队提出的 RoBERTa 预训练模型, 相比 BERT, 有更大的模型参数量、更大的训练批次数和更多的训练数据, 训练方法上 RoBERTa 去掉下一句预测任务并使用动态编码.

3.4 实验结果

如表 3 所示, 在 THUCNews-S 数据集上的实验结果显示, 与 BERT、FastText、TextRCNN 和 DPCNN 相比, RDAG 模型的 Acc 和 $F1$ 值均有不同程度的提升, 证实本文设计模型在中文文本分类任务的有效性. 但由于本文模型是根据自制中文金融领域长文本数据集特点而设计, 在通用领域的短文本分类任务上效果达不到最优, 但是超过绝大部分对比模型.

表 3 在 THUCNews-S 上的对比实验结果表 (%)

模型	*Acc*	*F*1
BERT	93.25	93.24
FastText	92.08	92.09
TextRCNN	92.91	92.87
DPCNN	90.76	90.76
EasyDL	**94.40**	**94.30**
RDAG	94.07	94.06

通过在 THUCNews-L 数据集上进行对比实验, 结果如表 4 所示, RDAG 模型相比于其他模型在准确率和 $F1$ 值都达到了最优, 说明在面对长文本分类任务时, 本文模型展现了较好的性能, 同时也证实本文根据自制数据集文本长度较长这一点特点设计的模型有效. 具体而言, RDAG 模型与 FastText 模型、RoBERTa-WWM+DPCNN 模型和 RoBERTa-WWM+BiGRU 模型相比, 其 $F1$ 值分别提升了 8.69%、1.65% 以及 0.94%. 这表明通过改进 DPCNN 模型中的等长卷积层, 并引入自注意力机制, 能有效提取到长文本数据的上下文依赖信息和词与词之间的关联信息, 再联合 BiGRU 模型, 能够进一步提高模模型特征提去能力. 在一定程度可以保证文本分类结果的准确率和标注质量.

表 4 在 THUCNews-L 上的对比实验结果表 (%)

模型	*Acc*	*F*1
FastText	89.04	88.87
EasyDL	96.62	95.20
RoBERTa-WWM	96.18	96.14
RoBERTa-WWM+DPCNN	95.92	95.91
RoBERTa-WWM+BiGRU	96.64	96.62
RDAG	**97.56**	**97.56**

通过在 THUCNews 数据集上的对比实验, 证明了本文所提模型在中文文本分类的有效性. 为了进一步证明本文模型在自制金融数据集上的优越性, 本文在自制金融数据集上做了对比消融实验. 从表 5 可以看出, 在该自制金融数据集上, RDAG 模型的得分明显高于其他比较模型. 这表明, 长距离依赖关系、深层次特征信息、上下文语义信息和词与词之间的关系对金融文本的分类性能有着重要的影响. 本文提出的 RDAG

模型充分考虑到以上因素, 尽可能提取并融合各类文本特征, 最终使得在金融数据集上的分类效果有一定程度提升. 与单一的抽取某一种特征信息的网络模型 FastText、TextRCNN 和 DPCNN 相比, 本模型准确率和 *F*1 值分别提高约 7% 和 10%. 直接使用 BERT 预训练模型接上 Softmax 分类器得到的 *Acc* 和 *F*1 值远低于本模型. 而 EasyDL 平台和 RoBERTa-WWM 模型虽然准确率不相上下, 但 *F*1 值相差近 11%.

表 5 在自制金融数据集上对比实验结果表 (%)

模型	*Acc*	*R*	*P*	*F*1
FastText	82.55	65.10	62.89	63.51
TextRCNN	84.27	68.18	69.51	68.31
DPCNN	85.61	66.60	72.59	68.43
EasyDL	86.30	65.00	63.70	64.10
RoBERTa-WWM	87.77	77.90	76.91	75.91
RoBERTa-WWM+DPCNN	89.93	76.65	78.95	77.36
RoBERTa-WWM+BiGRU	90.70	79.29	79.66	79.21
RDAG	**91.34**	**80.50**	**81.32**	**80.33**

3.5 消融实验

为证明模型中各个模块对整体效果的提升作用, 在自制金融数据集上进行了消融实验, 并得到各模型的 *Acc* 和 *F*1 值. 各模型实验结果如表 6 所示.

表 6 模型添加不同模块的对比实验结果表 (%)

模型	*Acc*	*F*1
BERT	76.18	75.33
BERT-DPCNN	76.50	75.42
RoBERTa-WWM	87.77	75.91
RoBERTa-WWM+DPCNN	89.93	77.36
RoBERTa-WWM+BiGRU	90.70	79.21
RoBERTa-WWM+DPCNN +BiGRU	91.08	80.10
RDAG	**91.34**	**80.33**

从表 6 得出: 通过对 BERT 和 RoBERTa-WWM 的实验结果进行分析可以看出, RoBERTa-WWM 预训练模型通过全词掩码结构和获取到的丰富词级别特征信息, 在自制金融长文本数据集上的分类效果明显优于 BERT 预训练模型. 加入 DPCNN 模型进一步提升文本特征信息的多样性, *Acc* 和 *F*1 值也明显上升, 验证了 DPCNN 模型对金融领域文本分类的积极影响. 对 RoBERTa-WWM、RoBERTa-WWM+BiGRU 和 RoBERTa-WWM+DPCNN+BiGRU 进行分析得到, 联合 DPCNN 模型与 BiGRU 模型能够得到文本的多元语义特征, 提高文本分类的准确率. 本文提出的 RDAG 模型与 RoBERTa-WWM+BiGRU 模型以及 RoBERTa-WWM+DPCNN+BiGRU 模型进行对比, 从实验结果可以看出, 引入自注意力机制和调整原始 DPCNN 模型的操作, 增强了 RDAG 模型对上下文语义信息和词与词之间的关联信息的捕捉能力, 有效地提升了文本分类的性能. 这主要是自注意力机制引入后对特征向量进行权重分配, 降低信息量少的特征向量权重值, 由此优化文本特征地表达能力, 达到进一步提高 RDAG 模型在金融领域长文本数据集上的分类准确率.

4 分析与讨论

通过第 3.4 节的 3 项实验表明, 不论是在公开数据集上还是在自制金融数据集上, 本文提出的 RDAG 模型对于中文文本分类都有着优异的分类效果, 尤其是针对长文本. 由于公开短文本数据集的文本长度的影响, RDAG 模型分类效果略低于百度提出的 EasyDL 模型. 在第 3.5 节消融实验中, 通过将各影响模块进行组合实验, 与 RDAG 模型效果对比得到, 改进的 DSACNN 模型相比于 DPCNN 模型, 抽取到更多词与词之间的特征信息和上下文长距离依赖关系, 使得 *Acc* 和 *F*1 值比优化前的 DPCNN 模型均提升约 0.3%. RDAG 模型中 DSACNN 模块和 BiGRU 模块联合抽取到文本的长距离依赖、词与词之间的关系、上下文语义信息和深层次特征, 使得实验结果明显优于其他对比模型, 且分类效果和稳定性更优.

5 结束语

本文提出了一种创新的金融长文本特征提取分类模型, 解决了金融长文本中词与词之间的关联信息、长距离依赖关系以及深层次语义特征信息提取问题, 为金融长文本分类问题提供新的解决方案. 并通过实验证明, 该模型与其他模型相比, 在金融长文本分类任务上更具有针对性和有效性. RDAG 模型通过削减 DPCNN 模型中卷积层的数量, 引入自注意力机制, 并将 BiGRU 模型作为辅助算法联合训练出针对金融文本特点的全新有效分类模型. 其中, 卷积层的改进保留更多文本特征, 自注意力机制提取单词之间的关系信息, 使模型变得简单、直接、易于使用.

然而, 由于该模型是基于金融文本数据集, 对于文本长度偏长的数据集, 其分类结果更好. 后续将会从文本长度和模型泛化能力的角度出发, 对模型进行优化和改进.

参考文献

1 Duan LG, Di P, Li AP. A new naive Bayes text classification algorithm. TELKOMNIKA: Indonesian Journal of Electrical Engineering, 2014, 12(2): 947–952.

2 Zhou Y, Li YW, Xia SX. An improved KNN text classification algorithm based on clustering. Journal of Computers, 2009, 4(3): 230–237.
3 Colas F, Brazdil P. Comparison of SVM and some older classification algorithms in text classification tasks. Proceedings of the 19th IFIP International Conference on Artificial Intelligence in Theory and Practice. Santiago: Springer, 2006. 169–178.
4 Gu JX, Wang ZH, Kuen J, *et al*. Recent advances in convolutional neural networks. Pattern Recognition, 2018, 77: 354–377. [doi: 10.1016/j.patcog.2017.10.013]
5 McClelland JL, Elman JL. The TRACE model of speech perception. Cognitive Psychology, 1986, 18(1): 1–86. [doi: 10.1016/0010-0285(86)90015-0]
6 Hochreiter S, Schmidhuber J. Long short-term memory. Neural Computation, 1997, 9(8): 1735–1780. [doi: 10.1162/neco.1997.9.8.1735]
7 Cho K, van Merriënboer B, Gulcehre C, *et al*. Learning phrase representations using RNN encoder-decoder for statistical machine translation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Doha: ACL, 2014. 1724–1734.
8 余本功, 张连彬. 基于 CP-CNN 的中文短文本分类研究. 计算机应用研究, 2018, 35(4): 1001–1004.
9 Zhou P, Qi ZY, Zheng SC, *et al*. Text classification improved by integrating bidirectional LSTM with two-dimensional max pooling. Proceedings of the 26th International Conference on Computational Linguistics: Technical Papers. Osaka: The COLING 2016 Organizing Committee, 2016. 3485–3495.
10 Zulqarnain M, Ghazali R, Ghouse M, *et al*. Efficient processing of GRU based on word embedding for text classification. JOIV: International Journal on Informatics Visualization, 2019, 4(3): 377–383.
11 金宁, 赵春江, 吴华瑞, 等. 基于 BiGRU_MulCNN 的农业问答问句分类技术研究. 农业机械学报, 2020, 51(5): 199–206.
12 张博, 孙逸, 李孟颖, 等. 基于迁移学习和集成学习的医学短文本分类. 山西大学学报 (自然科学版), 2020, 43(4): 947–954.
13 Dai BY, Li JL, Xu RY. Multiple positional self-attention network for text classification. Proceedings of the 34th AAAI Conference on Artificial Intelligence. New York: AAAI, 2020. 7610–7617.
14 张海丰, 曾诚, 潘列, 等. 结合 BERT 和特征投影网络的新闻主题文本分类方法. 计算机应用, 2022, 42(4): 1116–1124.
15 范昊, 何灏. 融合上下文特征和 BERT 词嵌入的新闻标题分类研究. 情报科学, 2022, 40(6): 90–97.
16 胥桂仙, 张子欣, 于绍娜, 等. 基于图卷积网络的藏文新闻文本分类. 数据分析与知识发现, 2023, 7(6): 73–85.
17 Mnih V, Heess N, Graves A, *et al*. Recurrent models of visual attention. Proceedings of the 27th International Conference on Neural Information Processing Systems. Montreal: MIT Press, 2014. 2204–2212.
18 Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate. Proceedings of the 3rd International Conference on Learning Representations. San Diego: ICLR, 2015.
19 杨兴锐, 赵寿为, 张如学, 等. 结合自注意力和残差的 BiLSTM_CNN 文本分类模型. 计算机工程与应用, 2022, 58(3): 172–180.
20 Vaswani A, Shazeer N, Parmar N. Attention is all you need. Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 6000–6010.
21 Devlin J, Chang MW, Lee K, *et al*. BERT: Pre-training of deep bidirectional Transformers for language understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis: ACL, 2019. 4171–4186.
22 Liu YH, Ott M, Goyal N, *et al*. RoBERTa: A robustly optimized BERT pretraining approach. arXiv:1907.11692, 2019.
23 Cui YM, Che WX, Liu T, *et al*. Pre-training with whole word masking for Chinese BERT. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021, 29: 3504–3514. [doi: 10.1109/TASLP.2021.3124365]
24 王仁超, 张毅伟, 毛三军. 水电工程施工安全隐患文本智能分类与知识挖掘. 水力发电学报, 2022, 41(11): 96–106.
25 Johnson R, Zhang T. Deep pyramid convolutional neural networks for text categorization. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Vancouver: ACL, 2017. 562–570.
26 加米拉·吾守尔, 吴迪, 王路路, 等. 基于多卷积核 DPCNN 的维吾尔语文本分类联合模型. 中文信息学报, 2021, 35(7): 63–71.
27 Yu SJ, Liu DL, Zhang Y, *et al*. DPTCN: A novel deep CNN model for short text classification. Journal of Intelligent & Fuzzy Systems, 2021, 41(6): 7093–7100.
28 Liu B, Zhou Y, Sun W. Character-level text classification via convolutional neural network and gated recurrent unit. International Journal of Machine Learning and Cybernetics, 2020, 11(8): 1939–1949. [doi: 10.1007/s13042-020-01084-9]
29 陈可嘉, 刘惠. 基于改进 BiGRU-CNN 的中文文本分类方法. 计算机工程, 2022, 48(5): 59–66, 73.
30 Bojanowski P, Grave E, Joulin A, *et al*. Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics. Cambridge: ACL, 2017. 135–146.
31 Lai SW, Xu LH, Liu K, *et al*. Recurrent convolutional neural networks for text classification. Proceedings of the 29th AAAI Conference on Artificial Intelligence. Austin: AAAI Press, 2015. 2267–2273.

(校对责编: 牛欣悦)