基于深度学习的文本分类系统关键技术研究与模型验证

汪少敏，杨迪，任华
（中国电信股份有限公司上海研究院，上海 200122）

摘　要：大数据时代，文本分类是文本数据挖掘和文本价值探索领域的重要工作。传统的文本分类系统存在特征提取能力弱、分类准确率不高的问题。相对于传统的文本分类技术，深度学习技术具有准确率高、特征提取有效等诸多优势，有必要将深度学习技术引入文本分类系统，以解决传统文本分类系统存在的问题。在分析传统文本分类系统的基础上，提出了基于深度学习的文本分类系统的体系架构和关键技术，同时对传统分类模型、TextCNN、CNN+LSTM 多种分类模型进行了验证比对。

关键词：深度学习；文本分类；分类模型；神经网络
中图分类号：TP393
文献标识码：A
doi: 10.11959/j.issn.1000−0801.2018301

Key technology research and model validation of text classification system based on deep learning

WANG Shaomin, YANG Di, REN Hua
Shanghai Research Institute of China Telecom Co., Ltd., Shanghai 200122, China

Abstract: Text classification is very important to text data mining and value exploration. The traditional text classification system has problems of weak feature extraction ability and low classification accuracy. Compared with the traditional text classification technology, deep learning technology has many advantages such as high accuracy and effective feature extraction. Therefore, it is necessary to apply deep learning technology to the text classification system to solve the problems of the traditional text classification system. The traditional text classification system was analyzed, and the architecture and key technologies of text classification system based on deep learning were proposed. Finally, several classification models were verified and compared, including the traditional classification model, TextCNN and CNN+LSTM.

Key words: deep learning, text classification, classification model, neural network

1　引言

目前，人工智能领域能处理的三大类信息格式主要为：文本、语音和图像[1]。其中，文本处理由于其庞大的数据源、相对成熟的处理技术和广泛的应用需求，备受关注。文本处理中最典型的场景为文本自动分类，即由计算机系统自动将文本数据归类到预设好的类别中。文本自动分类

收稿日期：2018−10−20；修回日期：2018−12−10

有着重要的应用，如客服工单的自动分类、提供个性化新闻[2]、用户意图和情绪分析[3-4]。

文本分类系统是利用机器学习技术来实现文本样本的自动归类。文本分类系统主要包含特征提取和有监督的机器学习两部分。传统的文本分类系统多采用特征加权技术进行文本特征提取，机器学习部分主要基于贝叶斯分类器、SVM（support vector machine，支持向量机）、随机森林等浅层机器学习模型。传统文本分类系统虽然文本训练集的训练速度快、资源要求低，但存在以下问题：一是准确率不够高，实验验证准确率仅为 70%左右[5]；二是由于特征提取采用特征加权方式，处理大规模文本数据时，存在特征提取后数据的高纬度高稀疏问题，造成对文本数据的特征表达能力弱，丢失文本数据的关键信息；三是特征提取过程较为复杂，需要人工参与定制，成本较高。传统文本分类系统存在的以上问题和缺陷，迫切需要解决和改进。由于深度学习技术具有准确率高、特征提取有效的特点[6]，因此将深度学习技术引入文本分类系统，以解决传统文本分类系统的问题。

本文针对传统文本分类系统存在的局限性，在分析了传统文本分类系统的基础上，探讨了基于深度学习的文本分类系统架构和关键技术，并实验验证对比了多个机器学习模型。基于深度学习的文本分类系统能有效解决传统文本分类技术中特征提取复杂、特征表达能力弱的问题，同时，通过实验数据，验证了深度学习模型有效提高了文本分类的准确率，从而实现更准确的文本自动分类。所以，基于深度学习的文本分类系统为解决文本自动分类需求，提供了新的技术方向和实现手段。

2 传统文本分类系统

文本分类是个有监督的机器学习过程，以已标注的文本集为基础，通过分类器找出文本类别与文本特征之间的关系，然后利用这个关系模型对新的文本进行类别判断[7]。分类的过程分为训练和测试两步。首先，将标记好类别的文本数据训练集输入系统，训练分类器模型；其次，将未标记类别的文本数据测试集输入系统，以验证系统的准确率。模型和系统通过多轮训练和调优，达到准确率要求后，根据业务场景需求，部署到实际生产环境中，对未分类的文本数据进行自动分类。

传统文本分类系统主要由特征提取和分类器模型两部分组成。特征提取将文本数据转换为反映样本类别特征的信息形式[8]，以便后续分类器模型处理识别。分类模型是有监督的机器学习模型，用以识别分类文本样本。传统文本分类系统一般采用词频加权的方式实现特征提取；同时，采用统计学基础的浅层机器学习模型，如朴素贝叶斯分类算法、SVM、最大熵和随机森林等。传统文本分类系统结构如图 1 所示。

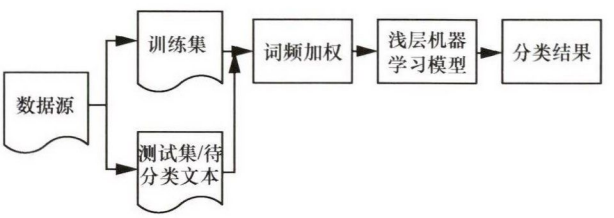

图 1 传统文本分类系统结构

词频加权对分词后的文本计算词语的权重，即计算某个词在样本中的重要程度[9]。计算完成后，得到每个词在样本中的权重列表。根据权重大小为样本提取特征词，作为浅层分类模型的输入。常用的词频加权方法为 TF-IDF。传统文本分类系统的特征提取方法存在以下缺陷：一是特征提取后的文本表示高维度高稀疏，不便于分类器处理；二是由于只统计了单个词的重要程度，丢失了词的上下文顺序；三是通过统计的方法对字符形式进行匹配分析，忽略了词的语义差异。以上缺陷造成传统文本分类系统的特征表达能力很弱，样本的特征信息丢失，影响分类结果。

传统文本分类系统采用的浅层机器学习模型由于结构简单，训练速度快，训练需要的计算资源较少。然而，浅层分类模型存在一定局限性，主要表现为对复杂函数的表示能力弱，造成针对复杂分类问题的泛化能力受到一定制约。所以，基于浅层分类模型的传统文本分类系统的相关分类能力受到限制，准确率有限。

3　基于深度学习的文本分类系统

由于传统文本分类系统存在的特征表达能力弱、分类模型处理复杂问题能力受限的问题，有必要引入先进的深度学习技术解决。基于深度学习的文本分类系统，针对传统文本分类技术存在的问题从多个方面做了改进和优化。

3.1　体系架构

基于深度学习的文本分类系统在实现时主要分为三大部分：文本预处理、词向量表示和深度学习模型，如图 2 所示。文本样本输入系统后，首先进行文本预处理，输出分好词的、更易于系统处理的文本数据；然后对这些数据进行词向量表示，进行特征提取，转换成深度学习模型能够学习训练和分析的格式；最后输入深度学习模型进行标记样本的学习或未标记样本的分类。

文本预处理是通过一系列手段，将文本数据提炼转换，去除信息价值低的文本数据，而最大可能地保留文本样本的特征信息，以节省后续的分析成本，提升分类效率和准确率。基于深度学习的文本分类系统的文本预处理主要包含 3 个部分：过滤非中文信息、模板提取和文本分词。

词向量表示是对文本预处理后的文本样本做进一步特征提取。词向量表示是基于深度学习的文本分类系统区别于传统文本分类系统的重要部分。不同于传统文本分类系统的词频加权方法，词向量表示利用神经网络学习的方法将文本数据转换为连续稠密的数据，不仅更便于深度学习模型分析，而且保留了文本的语义信息，从而解决了传统文本分析系统中，高纬度高稀疏的文本表示问题，以及忽略语义差异，从而导致特征提取弱的问题。

深度学习模型是文本分类系统的核心，它基于有监督的神经网络算法，对文本预处理后的样本进行学习训练或验证识别，从而输出文本自动分类的结果。由于深度学习模型的多层结构，提升了在训练过程中学习数据集特征的能力，能够实现复杂函数的逼近，所以深度学习模型解决了传统文本分类系统中分类模型对复杂问题泛化受限导致准确率不高的问题。

3.2　文本预处理

无论是用已标记文本样本训练深度学习模型，还是对未标记文本样本进行分析分类，首先都需要对文本数据进行预处理，包括过滤非中文信息、模板提取、文本分词和去停用词等操作。

3.2.1　过滤非中文信息

文本样本中包含数字、英文字母、标点、特殊符号、全角字符等非中文信息，这些信息对文本分类不起任何分类特征作用，所以在文本预处理过程中，应对文本样本中的数字、英文字母、标点、特殊符号等非中文信息进行过滤清除。

过滤文本样本中的非中文信息多采用正则表

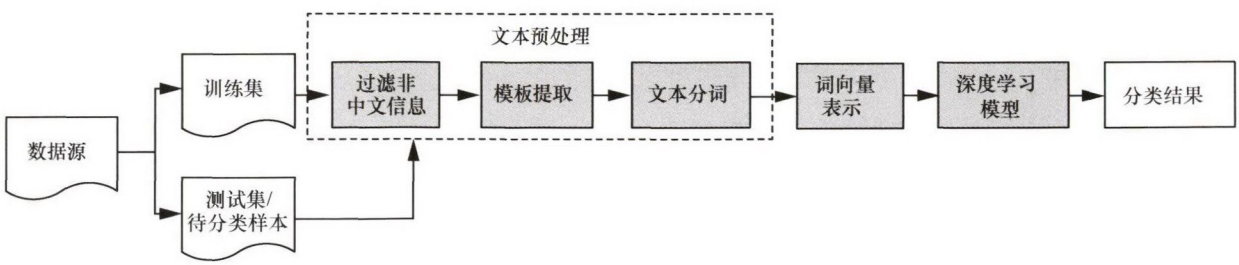

图 2　基于深度学习文本分类架构和流程步骤

达式的方法。如：“2017 年 09 月 10 日”“12.04 元人民币”“号码为 67099999988”等形式的信息，使用正则表达式过滤其中的数字字符。

3.2.2 模板提取

在一些应用场景中，文本样本往往遵循一定的模板格式，如：新闻报导的第一段是全篇的总结；客服人员按固定模板记录用户投诉工单。这些模板格式对文本样本的特征提取有重要作用。可以根据具体分类需求，通过模板提取文本样本中的关键信息，剔除其他信息，在后续处理中，只对提取的关键信息进行分析分类。例如，在客服记录的投诉工单自动分类应用中，客服记录投诉的模板见表 1。若分类需求为根据用户投诉的内容进行投诉类别分类，即可以根据客服记录模板，提取筛选出工单样本中的“受理内容”“问题描述”的文本，后续只对这些文本内容进行分析。

表 1 文本模板举例

模板 1	模板 2	模板 3
受理号码：	分类：	【客户号码】
受理内容：	现象：	【参与时间及活动】
核查情况：	故障地点：	【参与地点】
处理要求：	派单备注：	【问题描述】
回复时限：	昵称：	【预处理结果】
	IM 号：	
	渠道：	
	地区：	

通过模板提取文本样本中的关键信息，能够大大节省文本分类系统的计算时间和资源、提升系统效率，同时，排除非关键文本信息的干扰，提高分类准确率。

3.2.3 文本分词

要让机器理解识别文本数据，文本分词是文本预处理的重要步骤。后续的分类操作需要使用文本中的单词来表征文本，所以分词效果直接影响分类效果。文本分词的基本原理是分词算法根据预设的词典对样本进行匹配识别或标注训练。所以，文本分词包括两个部分：词典构造和分词算法。

词典包括通用语料词典和用户定义词典。常用的通用词典有搜狗语料库、国家语委语料库等。用户定义词典是通用词典未包括的、专业的单词。例如，电信运营商行业的应用场景中，用户定义词典包含天翼、翼支付、网关等。随着文本分类系统的使用，词典需要人工添加新词到词典中，适时定时地进行更新。

中文文本分词算法一般有字符串匹配分词和机器学习两大类实现方法，两种方法也可结合使用，如应用广泛的 Jieba 分词工具。字符串匹配分词的实现思路为，按照一定的策略匹配字符串和词典，以识别单词；机器学习分词的实现思路为：通过对大量汉字和单词进行标注训练，利用机器学习工具，识别文本中的词语。

3.2.4 去停用词

文本预处理时，为了节省后续的处理资源、提高效率，可以过滤某些对分类无作用的字或词，即“停用词”。停用词由人工整理生成，通过停用词表的形式输入文本分类系统中，一般在文本分词的步骤中调用相关函数或代码加载停用词表，实现停用词过滤。

停用词一般分为两类：

- 样本中随处可见的词，比如“工号”一词几乎在所有客服投诉工单中均会出现。这类词对文本分类没有信息作用，难以帮助分类，同时还会降低分析效率；
- 自身无明确意义的词，包括语气助词、副词、介词、连接词等，如“的”“在”之类。

3.3 词向量表示

文本预处理后，文本样本被转换为由单词组成的文本格式。文本预处理虽然完成了一部分特征提取的工作，但只是简单去除了样本中和分类无关的信息，并且这种高维度高稀疏的样本格式并不能很好地被深度学习模型进行识别处理。所

以，在输入深度学习模型之前，还需要对文本样本进行词向量表示，以完成文本样本的高质量特征提取和格式转换。

词向量表示的基本思路是通过机器学习训练将文本中的每个词映射成一个固定长度的向量。所有这些向量构成词向量空间，每个词的向量是该空间中的一点，可以根据词向量在该空间上的距离来判断词之间在语义上的相似性。词向量表示输出的是包含每个词向量的向量矩阵，即把文本数据从高维度高稀疏的编码方式，转换为和图像、语音类似的连续稠密数据。同时，词向量表示引用了语义信息，文本样本特征提取时，在一定程度上保留了文本样本的信息。所以，将词向量表示作为特征提取的手段，应用在文本分类系统中，能较好解决传统文本分类系统中词频加权方法造成的高纬度高稀疏数据、忽略语义信息，从而特征表达能力弱的问题。目前，应用较广的词向量表示工具为 Word2Vec。Word2Vec 是谷歌于 2013 年公开开源的词向量计算工具，它通过 CBOW（continuous bag-of-word）和 Skip-gram（continuous skip-gram gram）两种神经网络算法实现。

在基于深度学习的文本分类系统中，Word2Vec 词向量表示一般设为 50 维或 100 维，经过 Word2Vec 工具计算后，输出每个词的 50 维或 100 维向量。图 3 显示了在基于深度学习的文本系统中，Word2Vec 设为 50 维向量，一条文本样本在文本预处理后，进行 Word2Vec 计算，输出这条样本的词向量表示。

3.4　深度学习模型

词向量表示解决了文本特征提取的问题。基于深度学习的文本分类的另一个关键技术为深度学习模型。相对于传统文本分类系统，基于深度学习的文本分类系统用深度学习的神经网络模型替代浅层分类模型，作为分类器。由于深度学习的多层结构，提升了在训练过程中学习数据集特征的能力，能够实现复杂函数的逼近，所以，利用深度学习模型，可以解决传统文本分类系统中分类模型对复杂问题泛化受限导致准确率不高的问题。在第 4 节将验证比对浅层分类模型和深度学习模型针对文本分类的准确率。

深度学习最典型的模型为卷积神经网络（convolutional neural network，CNN）和递归神经网络（recurrent neural network，RNN）[10]，而适用于文本处理的神经网络则是利用 CNN 和 RNN 及其变体做进一步特征提取，以适应文本处理的需求，主要包括 LSTM、TextCNN、CNN+LSTM 等。

LSTM（long short term）网络是 RNN 的一种特殊类型，可以学习长期依赖信息。它利用 RNN 上下层的动态相关特性，学习文本中的上下文关系，同时通过网络结构的设计，解决了 RNN 的长

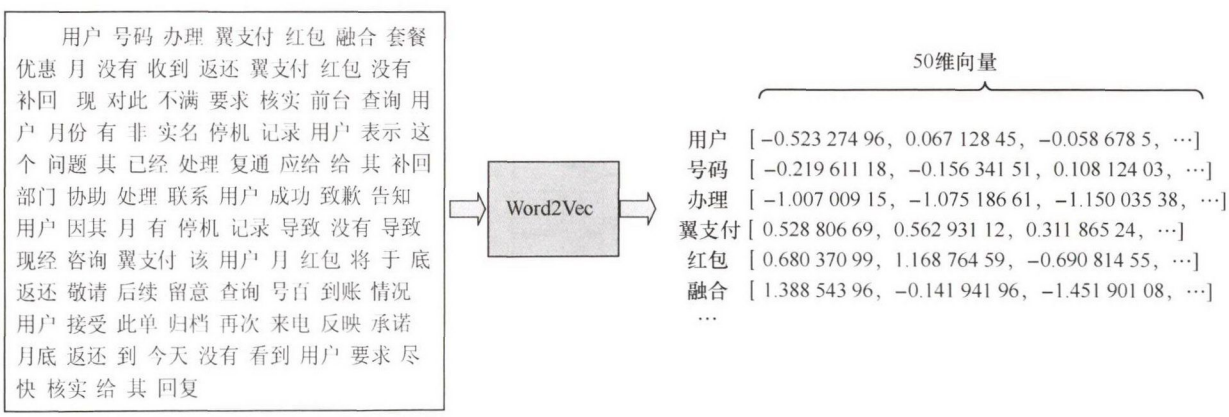

图 3　词向量表示示例

期依赖问题。TextCNN 是利用 CNN 来处理文本分类任务，它利用 CNN 捕捉局部相关性的特点，来提取句子中的关键信息，从而进一步进行文本样本的特征提取，实现更准确的分类。CNN+LSTM 的模型是结合 CNN 和 LSTM 网络，先使用 CNN 做局部特征提取，再用 LSTM 提取上下文关联信息。这样的网络结构结合了 CNN 和 LSTM 两种网络的优势，能够同时获取文本样本的特征表达和上下文信息。

深度学习分类模型的建立分为两步：一是用大量已标记类别的文本样本作为训练集训练分类类型；二是用去除标记类别的文本样本测试集验证模型的分类准确率。通过多轮训练和模型调优，实现文本分类模型的稳定。

以基于 TextCNN 的深度学习文本分类系统为例，具体网络结构和处理流程如图 4 所示。第一层是图 4 中最左边的 $n\times k$ 的词向量表示矩阵，n 是文本预处理和词向量表示后样本中的单词数，k 是词向量表示的维度，每行是样本中每个词的词向量。然后经过卷积层，第 3 层是 max pooling 层，在这一层不同长度的样本都变成定长的输出。最后一层是全连接的 softmax 层，输出该样本对应每个类别的概率，从而实现文本样本的分类。

4 模型验证和比对

为了验证基于深度学习的文本分类系统的可行性和优势，有必要对深度学习模型和传统文本分类模型进行验证和比对。

实验验证的数据集为中国电信翼支付业务 2017 年 8 月至 11 月期间的投诉工单。共有两个数据集，数据集 1 包含训练样本 4 109 条，测试样本 830 条，类别数设为 15；数据集 2 包含训练样本 3 191 条，测试样本 641 条，类别数设为 5。数据预处理基于 Python3/Regex/SKLearn 实现；模型训练和验证，基于 Python3/TensorFlow1.2 实现。实验分别验证了朴素贝叶斯、TextCNN、CNN+LSTM 3 种模型在两个数据集上的准确率。验证流程如图 5 所示。

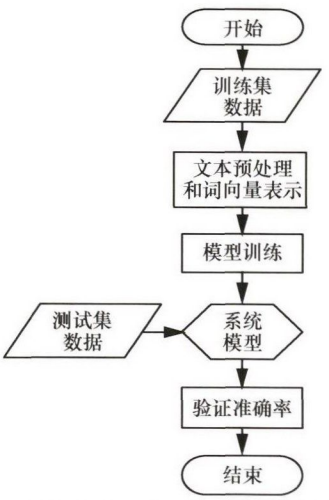

图 5 文本分类系统多个模型实验验证流程

验证结果见表 2 和图 6。从结果可以看出，基于 TextCNN、CNN+LSTM 和 Word2Vec 深度学习的文本分类系统，准确率要高于基于朴素贝叶斯

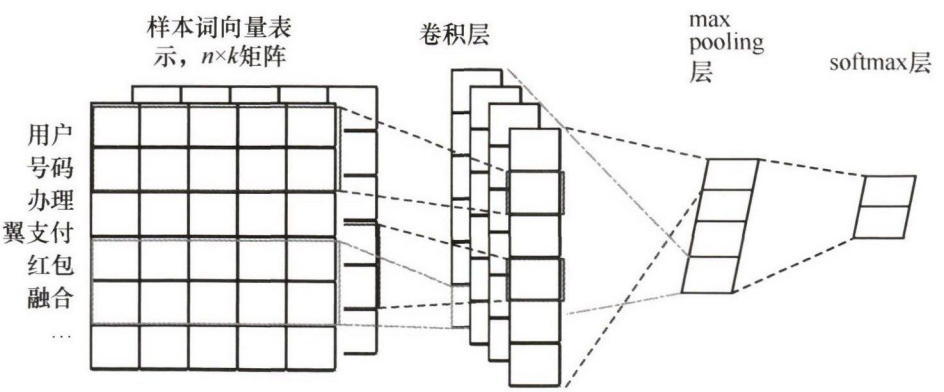

图 4 TextCNN 深度学习模型网络结构和处理流程

表 2　文本分类模型验证比对结果

分类模型	预处理	训练集样本数	测试集样本数	类别数	准确率
朴素贝叶斯	Jieba+TF-IDF	4 109	830	15	63.49%
TextCNN	Jieba+Word2Vec	4 109	830	15	67.89%
CNN+LSTM	Jieba+Word2Vec	4 109	830	15	68.64%
朴素贝叶斯	Jieba+TF-IDF	3 191	641	5	70.80%
TextCNN	Jieba+Word2Vec	3 191	641	5	73.81%
CNN+LSTM	Jieba+Word2Vec	3 191	641	5	75.24%

和 TF-IDF 的传统文本分类系统。这个验证结果和文中所分析的基于深度学习的文本分类系统优势相一致。

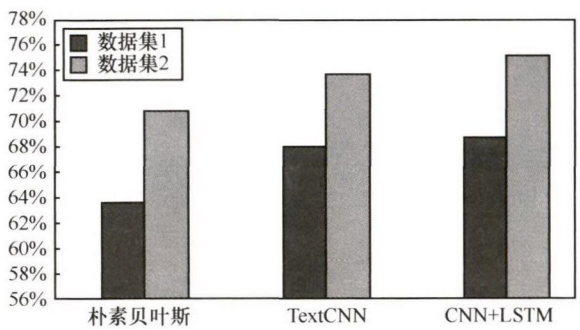

图 6　文本分类模型验证比对结果

5　结束语

文本自动分类一直有着广泛的应用需求，然而传统文本分类技术存在特征提取能力弱、准确率受限的问题。人工智能领域的深度学习技术为文本自动分类提供了新的技术方向，基于深度学习的文本分类系统较好地解决了传统文本分类系统的缺陷，有效提升了文本分类应用的准确率。

本文在分析了传统文本分类系统的基础上，将深度学习技术引入和应用到文本分类系统中，探讨了基于深度学习的文本分类系统的体系架构和关键技术，并进行了实验验证比对。由于时间和能力的限制，尚未对基于深度学习的文本分类系统的具体应用做进一步的研究，下一步将重点研究如何将深度学习的文本分类系统应用到不同需求场景中。

参考文献：

[1] 蔡自兴, 刘丽珏, 蔡竟峰, 等. 人工智能及其应用(第 5 版)[M]. 北京: 清华大学出版社, 2016.
CAI Z X, LIU L J, CAI J F, et al. Artificial intelligence and its application (fifth edition) [M]. Beijing: Tsinghua University Press, 2016.

[2] ANOTONELLIS I, BOURAS C, POULOPOULOS V. Personalized news categorization through scalable text classification[C]//Frontiers of WWW Research and Development-APWEB, Lecture Notes in Computer Science, January 16-18, 2006, Harbin, China. New York: ACM Press, 2006: 391-401.

[3] HU M, LIU B. Mining and summarizing customer review[C]//ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 22-25, 2004, Seattle, WA, USA. New York: ACM Press, 2004: 168-177.

[4] 蔡鑫, 娄京生. 基于 LSTM 深度学习模型的中国电信官方微博用户情绪分析[J]. 电信科学, 2017, 33(12): 136-141.
CAI X, LOU J S. Sentiment analysis of telecom official micro-blog users based on LSTM deep learning model[J]. Telecommunications Science, 2017, 33(12): 136-141.

[5] JOULIN A, GRAVE E, BOJANOWSKI P, et al. Bag of tricks for efficient text classification[J]. arXiv: 1607.01759v3[cs.CL], 2016.

[6] 余凯, 贾磊, 陈雨强, 等. 深度学习的昨天、今天和明天 [J]. 计算机研究与发展, 2013(9): 1799-1804.
YU K, JIA L, CHEN Y Q, et al. Deep learning: yesterday, today and tomorrow[J]. Journal of Computer Research and Develop-

ment, 2013(9): 1799-1804.

[7] 崔建明, 刘建明, 廖周宇. 基于 SVM 算法的文本分类技术研究[J]. 计算机仿真, 2013(2): 299-302.
CUI J M, LIU J M, LIAO Z Y. Research of text categorization based on support vector machine[J]. Computer Simulation, 2013(2): 299-302.

[8] TKACHENKO M, SIMANOVSKY A. Named entity recognition: Exploring features[Z]. 2012.

[9] ZHENG X Q, CHEN H Y, XU T Y, et al. Deep learning for Chinese word segmentation and POS tagging[C]// Empirical Methods in Natural Language Processing, Oct 18-21, 2013, Seattle, Washington, USA. [S.l.:s.n.], 2013: 647-657.

[10] LECUN Y, BENGIO Y, HINTON G. Deep learning[J]. Nature, 2015, 521(7553): 436-444.

[作者简介]

汪少敏（1983–），女，中国电信股份有限公司上海研究院高级工程师，主要研究方向为人工智能技术、自然语言处理、大数据和数据挖掘分析。

杨迪（1982–），男，中国电信股份有限公司上海研究院高级工程师，主要研究方向为人工智能技术、智能交互。

任华（1977–），女，中国电信股份有限公司上海研究院高级工程师，主要研究方向人工智能技术、智能客服。