基于上下文语义的新闻人名纠错方法

杨 越[1,2]，黄瑞章[1,2]，魏 琴[2*]，陈艳平[1,2]，秦永彬[1,2]

(1. 贵州大学计算机科学与技术学院 贵阳 550025；2. 贵州大学贵州省公共大数据实验室 贵阳 550025)

【摘要】新闻文本中的人名纠错存在以下难点：1) 人名中含有错误字段会影响甚至改变文本语义表达，故无法用传统命名实体识别方法识别句中人名；2) 人名字段的特殊性极易产生重名或者歧义，使得误报率增加，并提升了人名纠错的难度。为此，本文提出了一种基于上下文语义的新闻人名纠错方法。该方法使用卷积神经网络提取文本语义信息，并使用词激活力模型计算文本中其他词语与人名字段的关联程度来捕捉并使用文本上下文语义信息。同时，针对文本中人名字段中含有错误而导致的识别效果低下的问题，使用人名实体边界识别算法提高对文本中疑似含有错误人名的识别提取效果。实验结果表明，该方法能够有效地识别文本中的人名并对其中的错误内容进行纠正。

关 键 词 边界识别； 上下语义； 命名实体识别； 人名纠错

中图分类号 TP391.1 **文献标志码** A doi:10.3969/j.issn.1001-0548.2019.06.002

A News Name Correction Method Based on Context Semantics

YANG Yue[1,2], HUANG Rui-zhang[1,2], WEI Qin[2*], CHEN Yan-ping[1,2], and QIN Yong-bin[1,2]

(1. School of Computer Science and Technology, Guizhou University Guiyang 550025;

2. Public Big Data Laboratory of Guizhou, Guizhou University Guiyang 550025)

Abstract In news texts, incorrect fields in names will affect or even change the semantic expression of the text and the particularity of name fields will generate duplicate name or ambiguity. For solving these problems, this paper proposes a novel news name correction method based on context semantics. This method uses convolutional neural network to extract the semantic information of texts, and adopts word activation model to calculate the degree of association between other words and name fields in texts to capture and use the semantic information of text context. At the same time, aiming at the problem of low recognition caused by errors in the field of human name in texts, the entity boundary recognition algorithm of names is used to improve the recognition and extraction effect of names that are suspected to contain errors in the text. The experimental results show that the method can effectively identify the names in the text and correct the errors.

Key words boundary recognition; contextual semantics; named entity recognition; name error correction

在网络新闻中，人名表述错误较为常见，造成诸多不良影响。特别是时政类新闻，对人名的准确性要求很高，因此对文本中人名信息进行检查并纠错是一项重要的工作。

常见人名表述错误一般有两种情况：1) 拼写错误。在对目标人名进行输入时由于输入法拼写相似或者字体字型相似键入的错误人名；2) 语义错误。在编写文本时对文本语义不了解或是混淆，对人物的描述和其对应人名并不匹配，或者是人名实体字段对应属性表达错误。

针对第一种情况，通常使用编辑相似度方法来对文本进行纠错。但在实际应用场景中，单纯用编辑相似度来对疑似错误的人名进行纠错的效果并不理想，一是因为阀值不一定准确，会由于过于敏感或不敏感而纠错失败，另一个原因是使用存在错误字段的人名对其本身进行纠错会有很多不可控因素。而除了目标人名字段之外的句子上下文中含有比人名目标字段更多的语义信息，这些上下文信息能为人名纠错提供更多正确信息。

针对第二种情况，传统纠错方法无法应对语义错误，需要引入上下文语义信息，并根据语义关系判断识别错误信息来进行纠错。

收稿日期：2019－07－21；修回日期：2019－09－09

基金项目：国家自然科学基金联合基金重点项目(U1836205)；国家自然科学基金重大研究计划(91746116)；贵州省自然科学基金(黔科合基础[2018]1035)；黔科合重大专项字[2018]3002；贵州省重大应用基础研究项目(黔科合JZ字[2014]2001)；贵州省科技重大专项计划(黔科合重大专项字[2017]3002)

作者简介：杨越(1995－)，女，主要从事机器学习与自然语言处理方面的研究.

通讯作者：魏琴，E-mail: weiq@gzu.edu.cn

因此，本文提出一种基于上下文语义的新闻人名纠错方法，使用实体边界识别模型来识别句中疑似含有错误字符的人名字段，避免了因为目标字段中含有错误字符而导致的人名实体识别效果低下问题；同时利用卷积神经网络提取文本中人名所涉及的上下文语义，并加入词激活力模型计算文本中其他词语与人名字段的关联程度，从而得出该文本能否激活目标人名。

1 相关工作

对文本中的人名进行纠错，首先需要识别文本中的人名。文献[1-2]提出了LSTM-CRF模型来解决序列标注问题。文献[2]提出在英文NER任务中先使用LSTM来为每个单词由字母构造词并拼接到词向量再输入到LSTM[3]中，以捕捉单词前后缀等字母形态。文献[4]提出了在LSTM中加入基于词典的细胞，以提高针对特定实体的识别效果。但在实际的文本人名纠错应用场景中，由于人名中存在错误字符，影响文本的整体语义，使得人名识别效果并不理想。

针对实体消歧和实体对齐，文献[5]提出了一种提升文本中命中实体消歧鲁棒性的方法。文献[6]则结合社交网络的链接信息和聚类两种非监督框架对社交网络中的人名实体进行消歧。文献[7]提出了一种基于半监督协同训练的百科知识库实体对齐的方法，将实体对齐建模为一个带约束的二分类问题，使用半监督协同训练方法进行实体对齐。文献[8]则使用文本上下文依赖和句子语义进行事件线索检测。

上述方法都只考虑了某一个维度的特征，没有使用多个特征对其进行验证和特征综合。并且上述方法都没有提出一个针对文本中人名纠错的具体算法。

2 基于上下文语义的新闻人名纠错方法

本文提出一种基于上下文语义的新闻人名纠错方法。该方法先使用人名实体边界识别方法提取文本中含有疑似错误字符的人名字段，将提取到的文本信息分为人名字段字体拼写、文本语义相似程度、关键短语相关程度3个维度进行相似度计算和相似度整合约束，得到文本中人名相关信息的纠错结果。算法框架图如图1所示。

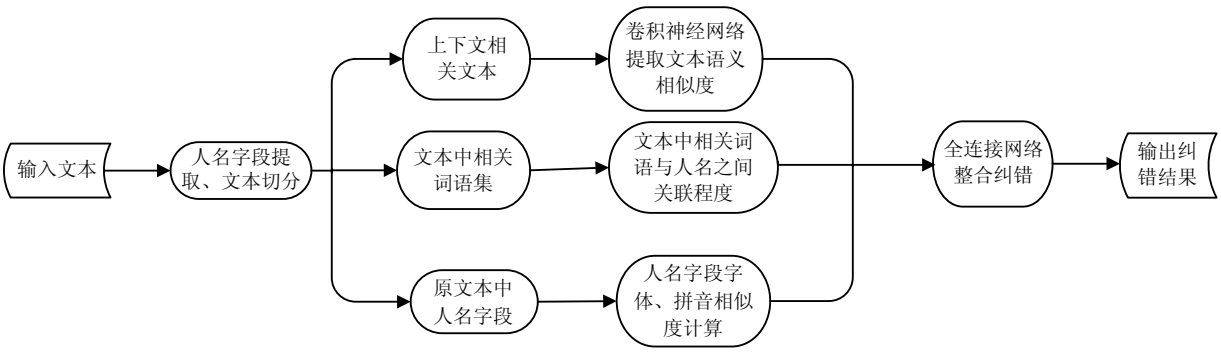

图1 基于上下文语义的新闻人名纠错方法框架图

2.1 基于字的左右双向实体边界识别模型

在对文本中人名进行纠错时，人名字段常常含有错误字符，若使用传统序列标注神经网络算法来对其进行识别，句中语义信息会受到错误字符影响从而降低含有错误字符的人名实体字段的识别效果。因此，本文提出并使用了左右双向实体边界识别模型，算法模型如图2所示。

本文使用了在现有方法中取得英文NER最好效果的模型方法[9-11]，使用LSTM模型作为神经网络的基础结构。对于句子级文本输入记作s，$s=c_1,c_2,\cdots,c_n$，c_i表示句子s中的第i个字符。对句子级文本输入每个字符标记的开始边界和结尾边界。

左右双向实体边界识别模型将句子级文本输入分为当前字符c_i，左子句$s_{\text{left}}=c_1,c_2,\cdots,c_{i-1}$，右子句$s_{\text{right}}=c_{i+1},c_{i+2},\cdots,c_n$。将左子句和右子句分别经过look-up层embedding映射后得到$\boldsymbol{X}_{\text{left}}$和$\boldsymbol{X}_{\text{right}}$，并通过LSTM模型得到左右子句的LSTM级输出$\boldsymbol{h}_{\text{left}}$和$\boldsymbol{h}_{\text{right}}$，将对于当前字符切分的左右子句输出进行拼合得到$\boldsymbol{H}=\left[\boldsymbol{h}_{\text{left}};\boldsymbol{h}_{\text{right}}\right]$，在最上层接入全连接层，将$\boldsymbol{H}$输入全连接层，得到当前字符$c_i$的最终标签，即当前字符为实体开始边界、结尾边界或不是实体边界。

2.2 文本中人名相关信息提取

2.2.1 卷积神经网络提取语义信息

本文使用卷积神经网络feature map形式来提取文本中上下文语义信息，得到文本语义与上文中提取到的人名的对应关系与相似度。

针对与目标人名字段相关的文本去掉文本中人

名字段，用卷积神经网络对字符序列进行处理，先对文本进行字符级embedding，将输入字符用低维稠密矩阵表示。用*w***h*大小的Filter在文本的embedding矩阵中提取文本中不可表达语义关系得到feature map特征数据，最后经过全连接层得到该上下文语义信息与该人名之间的相似程度[12]。

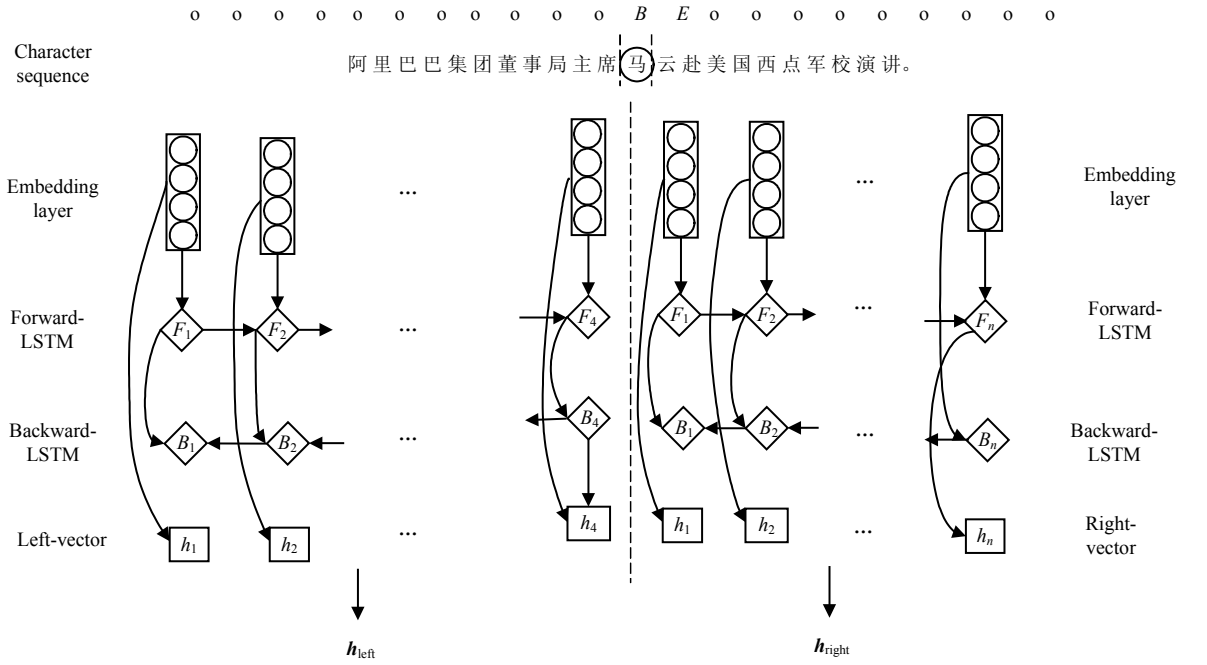

图2　人名边界识别模型，以“马”为模型当前字为例

其中，Filter的高(*h*)由不同范围内的词与词之间的相互关系来确定，可取为1或2。Filter的宽(*w*)一般都是词向量的维度，卷积后的feature map大小与卷积前输入数据的参数有关，它们满足以下关系[13]:

$$W_2 = \frac{(W_1 - F_W = 2P)}{S} + 1 \tag{1}$$

$$H_2 = \frac{(H_1 - F_h + 2P)}{S} + 1 \tag{2}$$

式中，W_2和H_2分别表示卷积后的feature map的宽度和高度；W_1和H_1代表卷积前句子矩阵的列和行；F_w和F_h分别表示filter的宽度和高度；而*P*和*S*则分别表示zero padding的数量和卷积过程中的步幅。池化层的主要作用为下采样，通过对feature map进行采样分析，过滤非重要成分，简化系统参数数量，从而提高运算效率，通过filter卷积后提取到最重要的特征参数，而后分别通过Dropout操作和全连接层进行分类[14]。

2.2.2　基于词激活力模型的词语相关联度

根据提取到的文本中人名字段，按目标人名字段前后词的词性，优先考虑名词，提取该词，与前文中提取到的人名字段组成词对，形如(马云，阿里巴巴)，(马云，双十一)。用词激活力模型判断文本中其他相关词语与人名相关字段的语义关系，文本中其他关键词语是否与目标人名字段有语义指向关系，或得到这些关键词语对应的语义关系的其他人名。

词激活力模型是一种全新的文本建模方法[15]，该方法旨在将词与词之间复杂的关系网络映射成计算机可读的词激活力矩阵，充分利用词语的上下文语义，更深层次地剖析出文本内在语义，建立符合文本的模型。

词激活力(word activate force, WAF)和词亲和力度(word affinity measure, WAM)是采用词关联度矩阵对文本深层语义 进行建模的两个关键概念。词激活力(WAF)表示一个词对另一个词的激发程度，即一个词的出现必然和其他的词有些潜在联系，例如，“马云”和“阿里巴巴”。对于给定词对(i, j)，称词语*i*为激活源，词语*j*为激活目标，词语*i*对词语*j*的词激活力定义如下：

$$\mathrm{waf}_{ij} = \frac{(f_{ij}/f_i)(f_{ij}/f_j)}{d_{ij}^2} \tag{3}$$

式中，f_i和f_j分别表示词语*i*和词语*j*在文本中出现的频次；f_{ij}表示词语*i*和词语*j*在设定共现距离内有序出现的频次；d_{ij}为两个词的平均共现距离。

词亲和度(WAM)用于表示在传统向量空间模型(VSM)中词语间缺失的潜在联系，其定义如下：

$$A_{ij}^{\text{waf}}=\left[\frac{1}{|K_{ij}|}\sum_{k\in K_{ij}}\text{OR}(\text{waf}_{ki},\text{waf}_{kj})\times \frac{1}{|L_{ij}|}\sum_{l\in L_{ij}}\text{OR}(\text{waf}_{il},\text{waf}_{jl})\right]^{1/2} \tag{4}$$

式中，K_{ij} 表示词对(*i*,*j*)的入链集合；L_{ij} 表示词对(*i*,*j*)的出链集合；OR(*x*,*y*)是一种重叠率计算函数。具体计算式如下：

$$K_{ij}=\{k\mid \text{waf}_{ki}>0 \text{ or } \text{waf}_{kj}>0\} \tag{5}$$

$$L_{ij}=\{l\mid \text{waf}_{il}>0 \text{ or } \text{waf}_{jl}>0\} \tag{6}$$

$$\text{OR}(x,y)=\frac{\min(x,y)}{\max(x,y)} \tag{7}$$

词亲和度 A_{ij}^{waf} 是词语*i*和词语*j*在词激活力矩阵中所有入链和出链重叠率的几何平均值，体现了两者在文本中的亲密程度。

针对在文本中提取得到的词对，经过词激活力模型计算，可以得到文本中其他词与目标人名字段之间的语义对应相关程度，该相关程度数值将作为词语相关联度特征进行下一步处理。

2.2.3 人名字段字体拼写相似度计算

本方法使用编辑距离来量化文本中提取到的疑似含有错误的人名字段和目标候选人名之间的拼写编辑相似度。编辑距离定义为针对2个字符串的差异程度的量化量测，量测方式是看至少需要多少次的处理才能将一个字符串变成另一个字符串。本方法中考虑人名字段的字体和拼音五笔输入法编辑距离，获取目标字段拼音及五笔输入法字符串，设置编辑距离阈值*W*，记文本中人名字段与候选人名字段之间编辑为*D*，则人名字段字体拼写相似度S_c计算为：

$$S_c=\frac{W-D}{W} \tag{8}$$

2.3 相似度整合

在具体文本语境中，特别是文本中含有不同类型错误的情况下，针对拼写相似度、语义相似度和词语关联程度这3方面特征各有侧重。人为给定特征权值占比或者根据固定公式计算等线性计算方式明显不能达到整合这些特征的效果，所以本文引入全连接层来进行相似度整合，全连接层网络结构如图3所示。

全连接层常用在卷积神经网络后对卷积池化后得到的Filter分类，而本文中经计算得到的3个特征可看做1×3维特征矩阵，用全连接层连接每个特征的相似度数值，以及BP反馈计算训练得到特征整合结果即该文本指向正确人名。

将拼写相似度、语义相似度、词语相关联程度3个特征数值作为全连接层输入，分别激活为 x_1,x_2,x_3，候选人名表示为 $N_1,N_2,\cdots,N_n$，整合进行得到纠错输出结果，如图3所示。

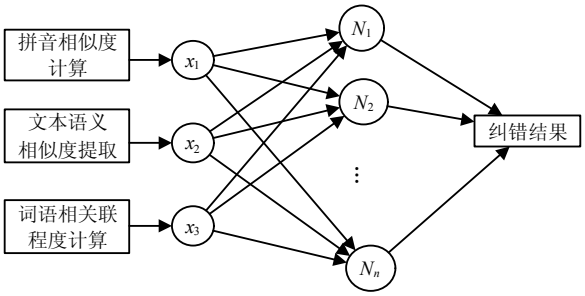

图3 全连接层网络模型

3 实验

3.1 实验数据

为了验证本文提出的基于深度学习的人名纠错方法的性能，从新浪网、人民网的新闻页面抓取共6 000篇时政类涉及共16个领导人的相关新闻作为模型训练集和测试集。对文档进行知识清洗，包括切词、停用词过滤、人名及其属性边界标注、语义标注等步骤。并人工标注加入人名及其属性相关错误负例句子级文本7 600句。以句子级文本对模型进行训练和测试，共57 000个句子，涉及16个人物和39个相关属性事件。

3.2 人名实体边界识别实验

进行人名实体边界实验计算时使用3.1节中提到人名属性数据集，对文本数据进行形如图2中所示标注。

本节采用准确率*P*(Precision)、召回率*R*(Recall)以及中和指标*F*值3项来进行评价。对于每个人名实体，*E*为使用本文中人名边界识别方法实验中得到的人名实体总数，E_1 是*E*中识别正确的人名实体数，E_2 是实验中涉及的人名实体总数。实验结果的准确率*P*、召回率*R*和*F*值分别为：

$$P=\frac{E_1}{E} \tag{9}$$

$$R=\frac{E_1}{E_2} \tag{10}$$

$$F=\frac{2PR}{R+P} \tag{11}$$

人名实体边界识别也是命名实体识别(NER)中的一项相关任务，本文使用BiLSTM-CRF、LSTM、CRF模型作为对比[1]。实验结果如表1中所示。

表1　人名识别实验对比效果

算法	*P*/%	*R*/%	*F*/%
本算法	93.12	92.79	93.11
CRF	86.62	86.98	86.53
LSTM	90.23	92.08	91.66
BiLSTM-CRF	91.79	92.08	91.66

实验结果表明，在句子级文本中人名实体含有疑似错误字段时，人名实体边界识别模型对人名实体识别的效果更好。

3.3　新闻人名纠错实验

进行人名及其属性纠错实验时，对算法模型的训练需要分为以下几步：1) 人名边界模型训练；2) 对卷积神经网络进行语义提取识别训练和词激活力模型数据收录计算；3) 利用拼写相似度、语义相似度、词语关联程度3个特征训练全连接层。

将文本中含有人名及其属性错误信息的数量，即模型应该识别错误信息并纠错的内容数量记为K，无错文本内容数量记为N。将需要纠错的内容正确纠错的数量记为TK，对文本中的错误内容未能识别到错误或者纠错错误的内容数量记为FK。对于原本就是正确的文本，模型判断该内容正确并未对其进行的内容数量记为TN，对正确文本进行纠错的内容数量记为FN。对纠错任务构造以下评价指标[16]，F_β为文本纠错效率，J_Δ为查准率，T_Δ为查全率：

$$\frac{1}{F_\beta}=\frac{2}{J_\Delta}+\frac{1}{T_\Delta} \tag{12}$$

其中：

$$J_\Delta=\frac{\mathrm{TK}}{\mathrm{TK}+\mathrm{FK}} \tag{13}$$

$$T_\Delta=\frac{\mathrm{TK}}{\mathrm{TK}+2\mathrm{FN}} \tag{14}$$

为了验证方法的效果，本文与字符级N-gram阀值替换模型[17]、利用HMM思想纠错生成候选项的方法[18]及利用SMT进行纠错的方法[19]进行对比。实验结果如表2所示。

表2　新闻人名纠错实验对比效果

算法	J_Δ/%	T_Δ/%	F_β/%
本算法	78.35	79.02	77.15
文献[17]的方法	66.27	66.27	66.35
文献[18]的方法	71.66	71.68	71.92
文献[19]的方法	70.25	70.78	71.72

实验结果表明，基于概率模型的传统纠错方法中，大多只考虑了基于N-gram模型上下字词出现的频率关系而很少考虑到文本间的语义关系，所以能识别的错误类型少，对未出现过的错误不能识别，纠错效果一般。而本文提出针对人名字段及其属性的纠错方法，从文本上下文语义和文本中词语和人名字段的相关联度两个方面较为准确地把握了文中的语义信息，减少了传统纠错方法对固定短语字词的依赖，所以本方法很大程度上增强了对文本的容错程度。同时，使用编辑距离作为一个特征也考虑了实际应用中由于拼写造成错误的情况，而使用全连接层对这3个特征进行整合，也平衡了各方面权值，增加了本方法对各种错误文本计算的稳定性。

4　结 束 语

本文提出了一种基于上下文语义的新闻人名纠错方法。使用卷积神经网络来提取深层语义信息，对于语义上的错误，由于句子级文本中含有错误信息，其本身含有的信息量对其进行纠错，不易发现其深层语义错误。同时，使用词激活力模型计算文本中其他词语与人名字段的相关联程度，从词的角度充分捕捉了文本上下文语义信息，并使用了左右双向边界识别模型来提高对文本中含有错误的人名字段识别的效果。实验结果表明，本文提出的基于深度学习的人名及其属性纠错方法能有效解决人名及其属性纠错问题。

参　考　文　献

[1] HUANG Z, XU W, YU K. Bidirectional LSTM-CRF models for sequence tagging[EB/OL]. (2015-05-06). https://arxiv.org/abs/ 1508.01991v1.

[2] LAMPLE G, BALLESTEROS M , SUBRAMANIAN S , et al. Neural architectures for named entity recognition [EB/OL]. (2016-04-07). https://arxiv.org/abs/1603.01360.

[3] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735-1780.

[4] ZHANG Y, YANG J. Chinese NER using lattice LSTM[EB/OL]. (2018-07-05). https://arxiv.org/pdf/1805.02023.pdf.

[5] HOFFART J , YOSEF M A , BORDINO I , et al. Robust disambiguation of named entities in text[C]//Conference on Empirical Methods in Natural Language Processing. [S.l.]: ACM, 2015: 782-792.

[6] BEKKERMAN R, MCCALLUM A. Disambiguating Web appearances of people in a social network[C]//International Conference on World Wide Web. [S.l.]: ACM, 2005: 463-470.

[7] 张伟莉, 黄廷磊, 梁霄. 基于半监督协同训练的百科知识库实体对齐[J]. 计算机与现代化, 2017(12): 92-97.
ZHANG Wei-li, HUANG Ting-lei, LIANG Xiao. Instance

alignment algorithm between encyclopedia based on semi-supervised co-training[J]. Computer and Modernization, 2017(12): 92-97.

[8] 王凯, 洪宇, 邱盈盈, 等. 融合上下文依赖和句子语义的事件线索检测研究[J]. 计算机科学与探索, 2018, 12(3): 423-431.
WANG Kai, HONG Yu, QIU Yin-yin, et al. Combining context dependency and sentence semantic representation for event nugget detection[J]. Journal of Frontiers of Computer Science & Technology, 2018, 12(3): 423-431.

[9] YAO Y, HUANG Z. Bi-directional LSTM recurrent neural network for Chinese word segmentation[C]//ICONIP 2016. [S.l.]: Springer, 2016: 345-353.

[10] MA X, HOVY E. End-to-end sequence labeling via Bi-directional LSTM-CNNs-CRF[EB/OL]. (2016-05-29). https://arxiv.org/abs/1603.01354.

[11] CHIU J P C, NICHOLS E. Named entity recognition with bidirectional LSTM-CNNs[EB/OL]. (2016-07-16). https://arxiv.org/abs/1511.08308.

[12] 周飞燕, 金林鹏, 董军. 卷积神经网络研究综述[J]. 计算机学报, 2017, 40(6): 1229-1251.
ZHOU Fei-yan, JIN Lin-peng, DONG Jun. Review of convolutional neural network[J]. Chinese Journal of Computers, 2017, 40(6): 1229-1251.

[13] KIM Y. Convolutional neural networks for sentence cl-assification[EB/OL]. (2014-09-03). https://arxiv.org/abs/1408.5882.

[14] 高彦琳, 战学刚, 迟呈英. 基于CNN-LSTM模型的情感分析研究[J]. 辽宁科技大学学报, 2018, 12(6): 469-474.
GAO Yan-lin, ZHAN Xue-gang, CHI Cheng-yin. Sentiment analysis based on CNN-LSTM model[J]. Journal of University of Science and Technology Liaoning, 2018, 12(6): 469-474.

[15] GUO J, GUO H, WANG Z. An activation force-based affinity measure for analyzing complex networks[J]. Scientific Reports, 2011, 1: 1-9.

[16] LIN C J, CHU W C. A study on Chinese spelling check using confusion sets and N-gram statistics[J]. International Journal of Computational Linguistics & Chinese Language Processing, 2015, 20(1): 23-27.

[17] CHIU H, WU J, JASON S. Chinese spelling checker based on statistical machine translation[C]//Proceedings of the 7th SIGHAN Workshop on Chinese Language Processing. [S.l]: ACL, 2013: 50-53.

[18] ZHANG S, XIONG J, HOU J, et al. HANSpeller++: A unified framework for chinese spelling correction[C]//Proceedings of the 8th SIGHAN Workshop on Chinese Language Processing. [S.l.]: ACL-IJCNLP, 2015: 38-45.

[19] LIU X, CHENG F, DUH K, et al. A hybrid ranking approach to Chinese spelling check[J]. ACM Transactions on Asian and Low-Resource Language Information Processing, 2015, 14(4): 1-17.

编 辑 蒋 晓