Software Engineering and Applications 软件工程与应用, 2023, 12(6), 915-921
Published Online December 2023 in Hans. https://www.hanspub.org/journal/sea
https://doi.org/10.12677/sea.2023.126089

Hans汉斯

基于BiLSTM-CRF的中文藏头诗敏感词检测算法

何亚楠，游福成

北京印刷学院信息工程学院，北京

收稿日期：2023年11月9日；录用日期：2023年12月19日；发布日期：2023年12月27日

摘　要

在数字化和社交媒体时代，藏头诗作为一种文化传承与现代表达相结合的文学形式，其内容监控成为了互联网平台管理的一个挑战。由于其特殊的构造方式，即每行的开头字连起来可以表达特定意义，这一特性使得其成为了隐藏敏感信息的一种手段。尤其是在社交媒体和即时通讯平台上，用户可能会利用藏头诗来规避敏感词过滤机制。本研究提出了一种基于双向长短期记忆网络(BiLSTM-CRF)的藏头诗敏感词检测算法。该算法首先采用词嵌入方法将文字表示成高维向量，再利用BiLSTM模型对藏头诗正反双向的上下文语义进行理解，并捕获文本序列中跨句藏头词的依赖关系，最后通过CRF模型根据标签相关性输出标记序列。我们对算法在不同类型的藏头诗数据集上进行了测试，结果显示该算法能够有效地识别出敏感词汇，具有较高的准确率和召回率。本算法对于监管自动生成的文本内容，尤其是在保护文化传承和遵守网络法规方面显示出其重要价值。

关键词

藏头诗，敏感词检测，BiLSTM-CRF

Chinese Hidden-Head Poem Sensitive Word Detection Algorithm Based on BiLSTM-CRF

Yanan He, Fucheng You

School of Information Engineering, Beijing Institute of Graphic Communication, Beijing

Received: Nov. 9th, 2023; accepted: Dec. 19th, 2023; published: Dec. 27th, 2023

Abstract

In the era of digitization and social media, acrostic poetry, as a literary form that combines cultur-

文章引用: 何亚楠, 游福成. 基于 BiLSTM-CRF 的中文藏头诗敏感词检测算法[J]. 软件工程与应用, 2023, 12(6): 915-921.
DOI: 10.12677/sea.2023.126089

al heritage with modern expression, has posed a challenge to internet platform management due to content monitoring. Because of its unique construction, where the initial letters of each line can convey a specific meaning when connected, this feature makes it a means of hiding sensitive information. Particularly on social media and instant messaging platforms, users may use acrostic poems to circumvent sensitive word filtering mechanisms. This study proposes a sensitive word detection algorithm for acrostic poetry based on Bidirectional Long Short-Term Memory Networks (BiLSTM-CRF). The algorithm first uses word embedding to represent the text as high-dimensional vectors, then utilizes the BiLSTM model to understand the semantic context of acrostic poems in both forward and backward directions and capture dependencies of acrostic words across sentences in the text sequence. Finally, the CRF model outputs label sequences based on label relevance. We tested the algorithm on various types of acrostic poetry datasets, and the results demonstrate that the algorithm can effectively identify sensitive words with high accuracy and recall. This algorithm has significant value for monitoring automatically generated text content, particularly in preserving cultural heritage and complying with internet regulations.

Keywords

Acrostic Poetry, Sensitive Word Detection, BiLSTM-CRF

1. 引言

藏头诗是一种汉语诗歌的形式，最早起源于中国，其特点是每一行的开头字连起来可以读出一个人名、地名、话题或某种意义的词语。然而，在现代社会，藏头诗也被用于隐藏敏感信息或不当内容。随着社交媒体和网络论坛的兴起，这种形式的敏感词汇检测成为了内容监管的一项挑战。由于藏头诗的结构特殊性，传统的敏感词检测算法往往无法准确地识别其中的敏感内容。因此，研究一种能够有效识别藏头诗中敏感词的算法显得尤为重要。

现在藏头诗已经演变成多种形式：藏头、藏尾、藏中、藏角、斜梯等，本文主要研究五种藏匿位置的五言藏头诗如图 1，目前缺少一种能够检测出藏头诗中隐匿敏感信息的方法。近些年在自然语言处理领域，基于深度学习的命名实体识别任务(Named Entity Recognition, NER)取得了重大突破，受此启发，本文提出了藏头诗敏感词检测的一种有效方法。本文通过大模型生成不同类型的藏头诗数据，将诗句中的敏感词作为实体进行标注构成实体识别数据集，提出了基于 BiLSTM-CRF 的藏头诗敏感词检测算法，通过在训练集上的学习，在测试集中表现出了很好的效果。

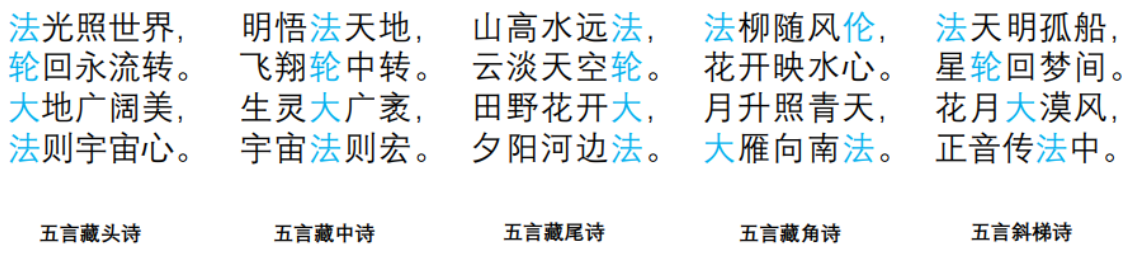

Figure 1. Five forms of five-character head-rhyme poems
图 1. 五种形式的五言藏头诗

2. 相关工作

传统的敏感词过滤算法较为简单，主要是单模式匹配和多模式匹配，这种基于字符串的匹配方法能够快速识别出文中敏感词[1] [2] [3]。该类方法实现简单，但是扩展性不强，需要动态更新敏感词表。为此有学者对 DFA 算法进行改进提出了 ST-DFA 算法[4]，当敏感词库更新时可以实时更新决策树，但该方法对变体敏感词检测的准确率不高。这些方法都只能检测出文本中连续文字组成的敏感词，不适用藏头诗中由跨句文字组成的敏感词，并且计算比较复杂，检测时间较长。

由于传统检测算法在敏感词及其变体上的局限性，有学者开始将敏感词检测任务与命名实体检测任务相结合进行研究。文献[5]就安全漏洞领域提出了一种命名实体识别方法，文献[6] [7]提出了基于 BiLSTM-CRF 的专业领域命名实体识别模型，文献[8]采用了基于 BERT-BiLSTM-CRF 的模型进行敏感词及其变体识别。通过这些研究，本文先利用大语言模型生成诗歌数据，然后人工标注敏感词构成数据集，在此数据集上构建命名实体识别模型，为藏头诗敏感词检测提出了有效的解决方案。

3. 本文方法

3.1. BiLSTM 模型

Hochreiter 提出 LSTM [9]是为了解决传统的 RNN 在处理长序列时梯度消失或梯度爆炸问题。它通过门控制机制(包括遗忘门、输入门、输出门)来调节信息的流动，能够较好地保存长期依赖信息。计算公式如下：

$$\begin{cases} f_t = \sigma\left(W_f \cdot \left[h_{t-1}, x_t\right] + b_f\right) \\ i_t = \sigma\left(W_i \cdot \left[h_{t-1}, x_t\right] + b_i\right) \\ \widetilde{C}_t = \tanh\left(W_C \cdot \left[h_{t-1}, x_t\right] + b_C\right) \\ C_t = f_t * C_{t-1} + i_t * \widetilde{C}_t \\ o_t = \sigma\left(W_o \cdot \left[h_{t-1}, x_t\right] + b_o\right) \\ h_t = o_t * \tanh\left(C_t\right) \end{cases} \tag{1}$$

式中： f_t, i_t, o_t 表示时间步t的遗忘门、输入门和输出门的激活向量； C_t 表示更新后的记忆元状态； h_t 为时间 t 的隐状态；的 W 为权重矩阵；b 为偏置量。

BiLSTM 即双向长短时记忆网络(Bidirectional Long Short-Term Memory)在标准的 LSTM 网络上进行了扩展，通过将数据正向和反向输入两个独立的 LSTM 网络，然后将它们的信息整合，以此来提高对上下文的理解能力。BiLSTM 的结构如图 2 所示。

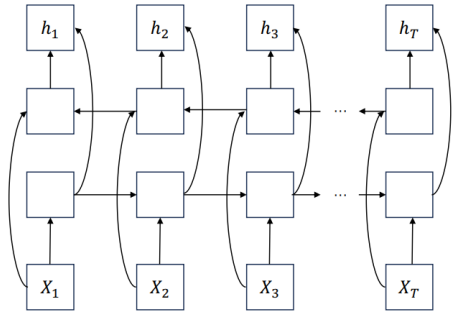

Figure 2. BiLSTM network structure
图 2. BiLSTM 网络结构

对于一个给定的序列 $X=(x_1,x_2,\ldots,x_T)$，BiLSTM 通过正向 LSTM 按顺序遍历序列，计算隐藏状态序列 $\overrightarrow{h_1}$，$\overrightarrow{h_2}$，…，$\overrightarrow{h_T}$。通过反向 LSTM 按逆序遍历序列，计算隐藏状态序列 $\overleftarrow{H_T}$，…，$\overleftarrow{H_2}$，$\overleftarrow{H_1}$。然后，对于序列中的每个时间步t，BiSLTM 的输出 h_t 是正向隐状态和反向隐状态的拼接。

$$h_t=\left[\overrightarrow{h_T};\overleftarrow{H_T}\right] \tag{2}$$

接下来，BiLSTM 的每个时间步输出h_t通常会通过全连接层，这些层可以学习如何基于 BiLSTM 层提取的特征来执行分类任务。在全连接层之后，会应用一个激活函数 softmax，将全连接层的输出转换成概率分布对标签进行预测。

3.2. CRF 模型

在本文中，我们采用条件随机场(CRF)模型来优化 BiLSTM 输出的实体识别标签，以获得巡检文本中的最佳实体标签序列。CRF 模型[10]通过考虑输入序列X和其对应的标签序列 Y 之间的关系，定义了一个条件概率分布 $P(Y|X)$，这里的X和Y都是随机变量序列，假设 X 和 Y 等长，并且用序列 $x=x_1,x_2,\ldots,x_n$ 和 $y=y_1,y_2,\ldots,y_n$ 分别代表输入和标签数据。在随机变量对 (X,Y) 上，CRF 模型的构建依赖于特定的局部特征向量 f 以及与之相对应的权重向量 λ。每一个局部特征可以是一个状态特征 $s(y',x,i)$，或者是一个转移特征 $t(y,y',x,i)$。y 和 y' 表示标签序列中的元素，x 表示输入的序列，而 i 指的是序列中的具体位置。

CRF 模型利用全局特征向量来表达输入序列 x 和相应标签序列 y 之间的关系。全局特征向量是位置信息 i 上所有局部特征函数 $f(y,x,i)$ 的累加即：

$$F(y,x)=\sum_i f(y,x,i) \tag{3}$$

基于此，CRF 模型定义的标签序列 y 给定输入序列 x 的条件概率分布可以表述为：

$$p_\lambda(Y|X)=\frac{\exp(\lambda F(Y,X))}{Z_\lambda(X)} \tag{4}$$

其中，$Z_\lambda(x)$ 是规范化因子，确保所有可能的上的概率总和为 1，通过下式计算：

$$Z_\lambda(x)=\sum_y\left|\exp(\lambda F(y,x))\right. \tag{5}$$

对于给定的输入序列 x，最有可能的标签序列$\hat{y}$可以通过最大化条件概率得到：

$$\hat{y}=\arg\max_y\ p_\lambda(y|x)=\arg\max_y\ \lambda F(y,x) \tag{6}$$

最后，通过维特比算法计算转移得分矩阵，可以找到具有最大条件概率的输出序列，实现标签序列的预测。

3.3. 敏感词检测模型

本文将藏头诗敏感词看成一种特殊的实体，利用命名实体识别技术可以有效识别出敏感词实体。对比传统的检测算法，该算法在处理规模和速度上均有明显优势。本文构建的 BiLSTM + CRF 模型如图 3 所示。模型大体分成三个部分：词嵌入模型、双向长短期记忆网络模型和条件随机场模型 CRF。首先采用词嵌入方法将文字表示成高维向量，再利用 BiLSTM 模型对藏头诗正反双向的上下文语义进行理解，并捕获文本序列中跨句藏头词的依赖关系，最后由 CRF 模型根据上下文序列标签相关性输出最终的标记序列。

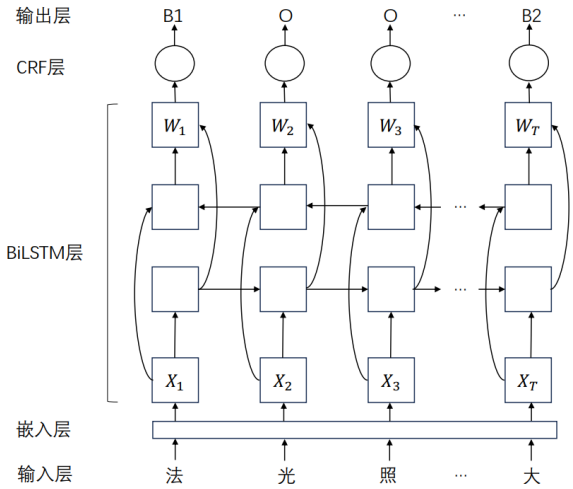

Figure 3. BiLSTM-CRF model framework
图 3. BiLSTM-CRF 模型框架

4. 实验

4.1. 数据集

由于藏头诗的敏感词检测属于新的研究方向，网络上没有现成的数据集。因为诗歌的复杂性很难靠自己撰写大量数据，而且含有敏感信息无法使用普通的诗歌生成工具。本文使用的语料库由大语言模型生成后经过人工筛选得到，然后通过程序对这些语料进行统一实体标注。因为诗句中的敏感词均由跨句文字构成，所以没有采用传统的命名实体标注体系，而是采用 B1-B4 的标签标注敏感词，用 O 标注非敏感无关字符，最终得到的数据集如图 4 所示。图中，标签 B1 代表敏感词的第一个字，B2 代表敏感词的第二个字依此类推。数据集共包含 1450 首诗，将数据集按照 8:1:1 的比例划分为训练集、测试集和验证集。

B1 O O O O O B2 O O O O O B3 O O O O O B4 O O O O O
法 光 照 世 界 ， 轮 回 永 流 转 。 大 地 广 阔 美 ， 法 则 宇 宙 心 。

Figure 4. Data set labeling format
图 4. 数据集标注格式

4.2. 实验环境与评价指标

本文的实验环境为：linux 操作系统、Python3.9、Pytorch 1.11.0。服务器 CPU 为 Intel(R) Core(TM) i7-12700H，内存 32 GB，GPU 为 RTX3070Ti。

为验证使用 BiLSTM-CRF 模型对敏感词识别的可行性和准确性，本文采用召回率 R、准确率 P 和 $F1$ 得分来评判模型的性能，各评价指标的计算方法如下：

$$P = \frac{TP + TN}{TP + TN + FP + FN} \tag{7}$$

$$R = \frac{TP}{TP + FN} \tag{8}$$

$$F1 = 2 \times \frac{\mathrm{Pr}ecision \times \mathrm{Re}call}{\mathrm{Pr}ecision + \mathrm{Re}call} \tag{9}$$

其中 TP (True Positives)是正确预测的正类数目，TN (True Negatives)是正确预测的负类数目，FP (False Positives)是错误预测为正类的负类数目，FN (False Negatives)是错误预测为负类的正类数目。

4.3. 实验结果

本文旨在评估 BiLSTM-CRF 模型在中文藏头诗中敏感词检测的性能。为了提升模型效果，反复训练最终将主要参数设置为：lstm_lr = 0.001，crf_lr = 0.1，epoches = 20，batch_size = 50。模型训练过程采用交叉熵损失函数和 Adam 优化函数。

根据上面设定的参数，模型的实验结果如表 1 所示。为了证明 CRF 的效果，本文在相同的数据集上分别训练了 BiLSTM 和 BiLSTM-CRF 模型。从结果来看 BiLSTM-CRF 的评估指标明显优于 BiLSTM，这是因为 BiLSTM 虽然能捕获敏感词的跨句依赖关系，但是无法学到输出标签的约束条件。所以 BiLSTM-CRF 在处理藏头诗的上下文信息方面具有更强的能力，能更有效地识别和定位敏感词。

Table1. Comparison of different model results
表 1. 不同模型结果比较

模型	Precision	Recall	F1
BiLSTM	88.22	95.52	91.72
BiLSTM-CRF	94.61	96.90	95.74

5. 结论

本文提出了一种 BiLSTM-CRF 模型，实现了对中文藏头诗中敏感词的检测。该方法是第一个将命名实体识别技术应用到中文藏头诗文本的敏感词检测领域，BiLSTM 模型能根据目标实体自动提取文本序列特征，CRF 模型可以学习到输出的标签之间的约束条件和依赖关系，实验展示了该模型在中文藏头诗敏感词检测任务中的准确性和有效性。

总体而言，本文为中文藏头诗的敏感词检测提供了一个强有力的工具，对于维护网络环境的健康与安全具有重要意义。随着文本形式的不断增加，藏匿敏感信息的方式越来越多而且更加隐晦难以发现，今后将继续改进模型和扩充语料库，逐步提升模型识别能力。

参考文献

[1] Sara Sood, Judd Antin, Elizabeth Churchill. (2012) Profanity Use in Online Communities. CHI '12: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 05-10 May 2012, New York, 1481-1490. https://doi.org/10.1145/2207676.2208610

[2] Liu, C., Wang, W.Y., Wang, M., *et al*. (2017) An Efficient Instance Selection Algorithm to Reconstruct Training Set for Support Vector Machine. *Knowledge-Based Systems*, **116**, 58-73. https://doi.org/10.1016/j.knosys.2016.10.031

[3] Guan, D.H., Yuan, W.W., Lee, Y.K., *et al*. (2008) Improving Supervised Learning Performance by Using Fuzzy Clustering Method to Select Training Data. *Journal of Intelligent & Fuzzy Systems*, **19**, 321-334.

[4] Xue, P.Q., Nurbol, and Wushour, I. (2016) Sensitive Information Filtering Algorithm Based on Text Information Network. *Computer Engineering & Design*, **37**, 2447-2452.

[5] 张若彬, 刘嘉勇, 何祥. 基于 BLSTM-CRF 模型的安全漏洞领域命名实体识别[J]. 四川大学学报(自然科学版), 2019, 56(3): 469-475.

[6] 黄炜, 黄建桥, 李岳峰. 基于 BiLSTM-CRF 的涉恐信息实体识别模型研究[J]. 情报杂志, 2019, 38(12): 149-156.

[7] 尤丽珏, 尹远芳. 基于 BiLSTM-CRF 模型的医学影像检查报告信息实体识别[J]. 微型电脑应用, 2023, 39(10): 134-137.

[8] 郑贤茹, 李柏岩, 冯珍妮, 等. 基于 BERT-BiLSTM-CRF 的网络敏感词及变体实体识别[J]. 计算机与数字工程, 2023, 51(7): 1585-1589.

[9] Dou, G., Zhao, K., Guo, M., *et al*. (2023) Memristor-Based LSTM Network for Text Classification. *Fractals*, **31**, Article ID: 2340040. https://doi.org/10.1142/S0218348X23400406

[10] 刘雪梅，程彭圣男，李海瑞，等. 基于字词向量的 BiLSTM-CRF 水利工程巡检文本实体识别模型[J/OL]. 华北水利水电大学学报(自然科学版), 1-9. http://kns.cnki.net/kcms/detail/41.1432.tv.20231102.1649.002.html, 2023-11-09.