Software Engineering and Applications 软件工程与应用, 2022, 11(6), 1430-1438
Published Online December 2022 in Hans. http://www.hanspub.org/journal/sea
https://doi.org/10.12677/sea.2022.116147

Hans汉斯

基于Bert-BiLSTM-CRF的标讯信息提取实现

李正军，涂著刚

贵阳高新数通信息有限公司，贵州 贵阳

收稿日期：2022年11月10日；录用日期：2022年12月22日；发布日期：2022年12月31日

摘　要

面对海量的标讯信息规模及复杂的数据结构，如何高效地挖掘潜在的数据价值，是能否有效实现招投标领域大数据应用的关键。本文通过大量数据标注，借助Bert-BiLSTM-CRF机器学习算法，对标讯信息的关键字段实现自动提取，有效实现标讯信息的结构化和价值化。

关键词

Bert-BiLSTM-CRF，数据价值，命名实体识别，深度学习，数据标注

Implementation of Bids Information Extraction Based on Bert BiLSTM-CRF

Zhengjun Li, Zhugang Tu

Guiyang Hi-Tech Data Communication Co., Ltd., Guiyang Guizhou

Received: Nov. 10th, 2022; accepted: Dec. 22nd, 2022; published: Dec. 31st, 2022

Abstract

In the face of massive scale and complex data structure of bidding information, how to efficiently tap the potential data value is the key to effectively implement big data applications in the bidding field. In this paper, with the help of Bert BiLSTM-CRF machine learning algorithm, the key fields of the banner information are automatically extracted through a large number of data annotations, effectively realizing the structure and value of the banner information.

Keywords

Bert-BiLSTM-CRF, Data Value, Named Entity Recognition, Deep Learning, Data Annotation

文章引用: 李正军, 涂著刚. 基于 Bert-BiLSTM-CRF 的标讯信息提取实现[J]. 软件工程与应用, 2022, 11(6): 1430-1438.
DOI: 10.12677/sea.2022.116147

1. 引言

政府采购及工程招投标领域数据的进一步开放，对数据的获取和综合利用能力将在未来成为衡量整个招投标大数据产业综合竞争力的新标杆。但面对海量的数据规模、复杂多样的数据结构、隐藏化的数据价值，如何在高效完成数据采集的同时，及时实现对信息的进一步抽取、识别、分析、建模及应用，挖掘出数据的隐藏商业价值，是标讯信息价值化路径的关键。

中文命名实体识别的实现算法是如今最为广泛使用的基于深度学习的方法。其中，前馈神经网络以单向方式进行信息传递，从而使得网络结构更加容易计算，但也削弱了神经网络的表达能力。循环神经网络通过对神经元接收自身的信息，而形成一个环路的网络结构，从而可以实现对文本序列中下文位置信息的记忆与处理。由于 RNN 中存在梯度爆炸或消失现象，从而导致神经元的短期记忆，存在长期依赖问题。LSTM 神经网络中，记忆单元通过捕捉到关键信息并将其在一定的时间内存储保存，从而解决梯度爆炸或消失问题。由于信息的输出将受到历史信息和后续信息两个方面影响，因此引入双向长短期记忆神经网络(Bidirectional LSTM, BiLSTM)。

本文基于 Bert + BiLSTM + CRF [1]的中文实体识别深度学习算法，通过分类集中标注，学习标注数据集中的语义信息以及一些规律来识别蕴含标讯信息中的关键字段信息，提升多种字段识别准确率，解决特定环境结合上下文实体识别问题，为构建完整有效的集数据采集、数据清洗、数据聚合、数据建模、数据产品化为一体的公开大数据产品奠定基础。

2. 算法

命名实体识别任务作为信息抽取的一项基础性工作，主要是通过序列标注的方法进行解决。通过使用神经网络模型处理命名实体识别的问题已经成为一种趋势，Hammerton 等人于 2003 年首次通过使用单向 LSTM 来处理命名实体识别问题，Collobert 等人在 2011 年将 CNN-CRF 模型运用到 NER 任务中，Lample 等人在 2016 年使用的 LSTM-CRF 模型在英文命名实体识别中取得突出的性能，成为了当前在命名实体识别问题中的主流模型之一[2]，CRF 的优点是能对隐含状态建模，学习状态序列的特点，通过在标签之间增加约束性的条件[3]，可以更好的符合语言逻辑的正确性，最终生成符合人类的语言模型。单词向量序列用作模型的输入表示，输入到双向 LSTM，通过训练提取特征信息，输出到 CRF 层，最后计算最佳注释序列。

2.1. BiLSTM

长短期记忆(LSTM)主要用于解决长序列训练中梯度消失和梯度爆炸的问题[4]。简单来说，就是相比普通的 RNN，LSTM 能够在更长的序列中有更好的表现。其主要结构如图 1 所示。

BiLSTM 模型添加了 LSTM 网络的后向层，前向和后向隐藏层一起形成 BiLSTM 网络模型的隐藏层，并最终加入输入层进行连接。对于输入层中句子的标签信息，前向 LSTM 用于从前往后表示句子的信息，后向 LSTM 则用于从后往前表示语句的信息。最后，上下文信息的连接是单词的最终表示。在输入层，句子中每个单词对应的单词向量可以输入到 BiLSTM 层，正向 LSTM 获得单词向量的正向隐藏层序列。

2.2. CRF

条件随机场(Conditional Random Field，简称 CRF)是一种判别式无向图模型[5]，生成式模型是直接对

联合分布进行建模，而判别式模型则是对条件分布进行建模，条件随机场则是判别式模型。跟隐马尔可夫模型通过联合分布进行建模不同，条件随机场试图对多个变量在给定观测值后的条件概率进行建模。通过上下文的分析，条件随机场分词会提升到更高的精度。但因为复杂度比较高，条件随机场一般训练代价都比较大。

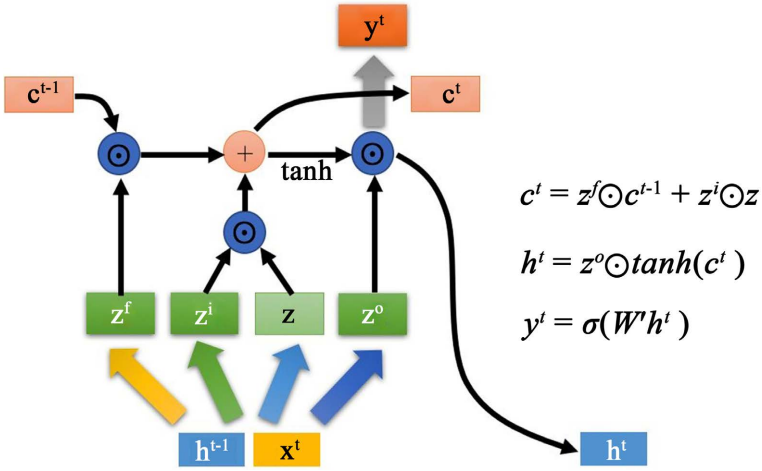

Figure 1. LSTM structure
图 1. LSTM 结构

2.3. Bert

BERT 的创新点在于它将双向 Transformer 用于语言模型，双向训练的语言模型对语境的理解会比单向的语言模型更深刻，Transformer 的原型包括两个独立的机制，一个 encoder 负责接收文本作为输入，一个 decoder 负责预测任务的结果[6]。BERT 的目标是生成语言模型，所以只需要 encoder 机制。Transformer 的 encoder 是一次性读取整个文本序列，而不是从左到右或从右到左地按顺序读取。

3. 实现流程

具有复杂结构的标讯的字段化、结构化是抽取、数据标注、模型迭代更新相互结合的一个流程，如图 2 所示。

3.1. 数据清洗

数据分析、定义数据清洗的策略和规则、搜寻并确定错误实例、纠正发现的错误以及干净数据回流。

3.2. 数据标注

采用联合标注：对一串连续的字标注相同的标签。在 NER 任务中，实体由一个或多个字组成，所以它属于联合标注任务。但是在联合标注中，相邻词语标签之间可能会存在依赖关系。这一问题可以通过标签转化的方式，把联合标注转化成原始标注解决。

3.2.1. 数据集 1

数量：数据量 8388 篇文本，16 字段。

标注方式和类型：对全文进行标注，

标注字段：

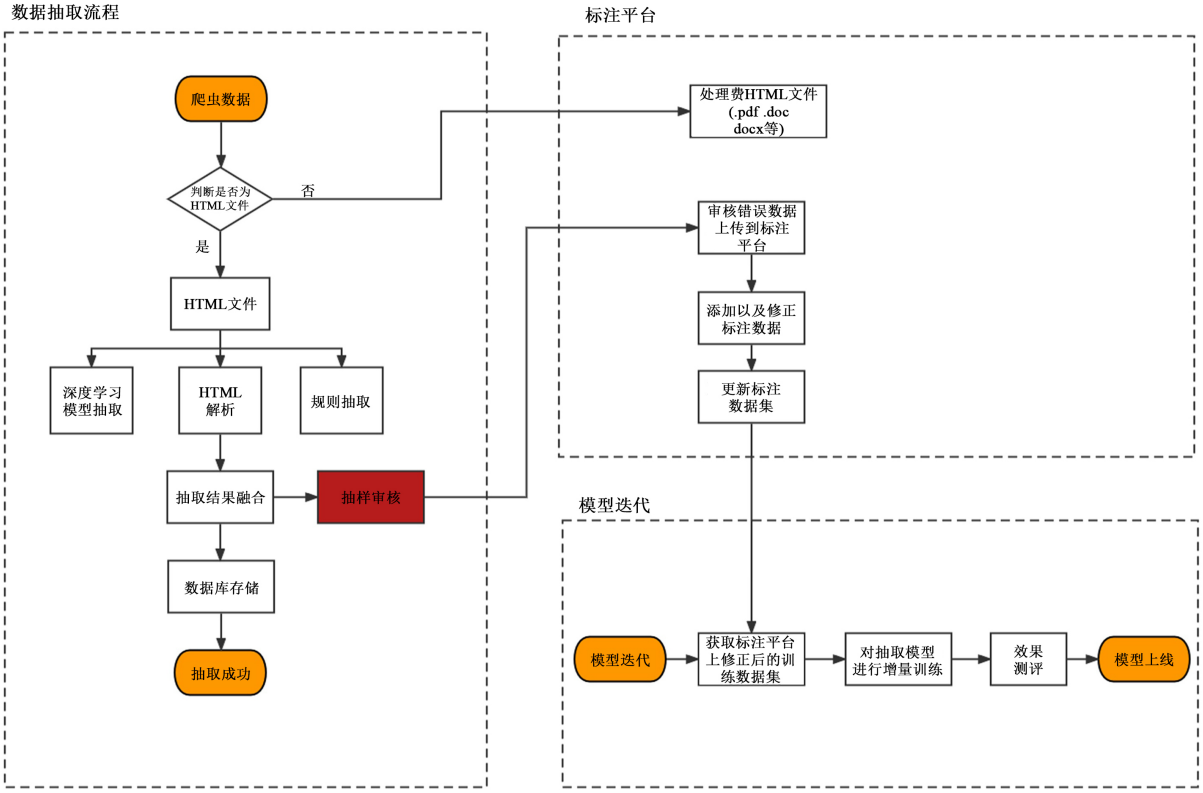

Figure 2. Information model identification process
图 2. 信息模型识别流程

项目名称、项目编号、项目预算、项目预算单位、报名截止时间、投标截止时间、业主、业主联系人、业主电话、代理机构、代理联系人、代理电话、中标机构、中标金额、中标金额单位、中标候选人

说明：该版本没有对数据进行分类标注抽取，全部类型数据为同一类进行 16 个字段的抽取。

3.2.2. 数据集 2

数量：33,000 左右，9 类，全部 34 字段。

标注方式和类型：对全文进行标注，

标注字段：

在后续的第二版本招标过程字段抽取中新增了其他字段，在招投标过程中，我们将招标过程中的文档分为预告、招标公告、采购公告、开标公告、评标公告、中标公告、采购结果、合同、其他 9 类。

3.3. 模型训练

对标注好的数据集使用 BERT + BiLSTM + CRF 模型进行训练。

1) 该模型训练过程中的 loss (训练损失)随着模型迭代次数的变化图如图 3 所示，该图反应了模型训练过程中损失下降的过程，对于一个模型损失越小越好。

2) 该模型训练过程中的验证集的 F1 值、precision (精确度)、recall (召回率)随着模型迭代次数的变化图如下所示，该图反应了模型在不断迭代学习过程中对验证数据集的一个测试，其中 F1、precision (精确度)、recall (召回率)值是衡量该模型好坏的标准，这三个值越大越好，其中 F1 值是通过 precision (精确度)、recall (召回率)计算得到的，F1 常用来表示一个模型整体好坏。图 4 展示了模型的准确率、召回率及 F 值的得分数随着迭代次数的变化情况。从图中看到，在开始迭代 20 次后，模型的性能结果逐步收敛，经过

20 次左右的迭代次数之后，模型评价指标分数值基本保持在一个较为稳定的数值，并处于一个很小范围浮动的状态。出于时间成本及性能结果的考虑，本文最终选取迭代次数为 60 的训练参数。

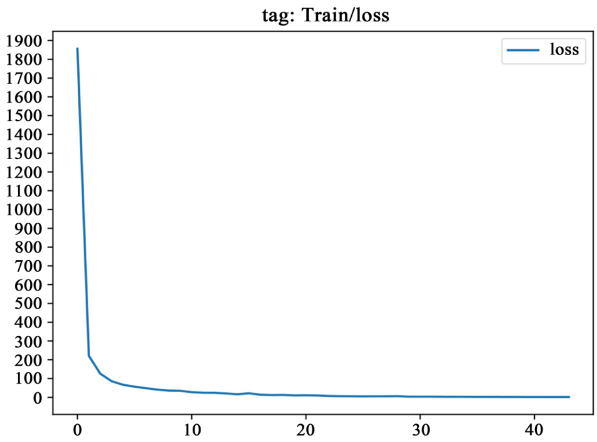

Figure 3. Loss (training loss) change chart
图 3. Loss (训练损失)变化图

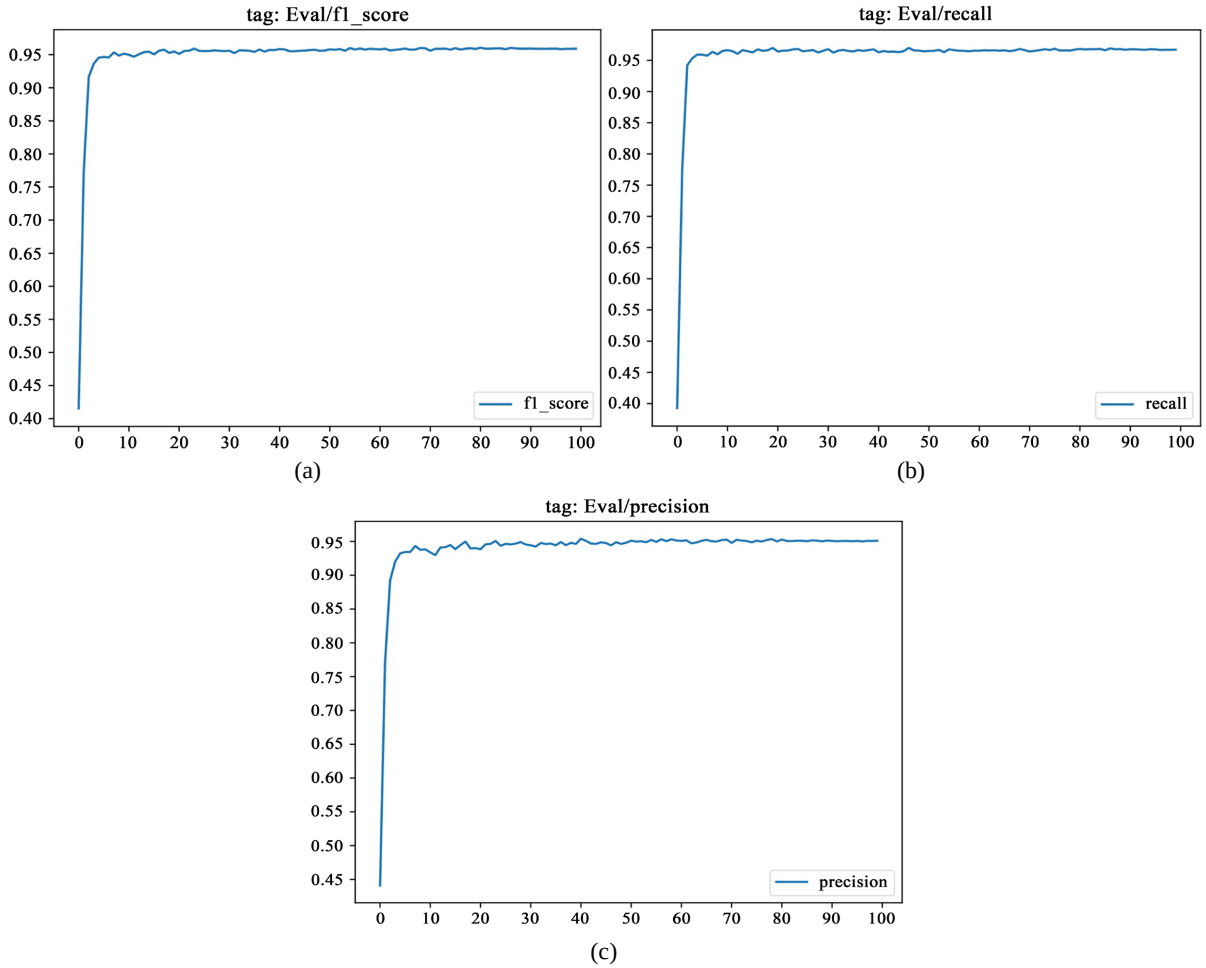

Figure 4. F1 value, precision and recall change chart
图 4. F1 值、精确度、召回率变化图

3) 训练步骤：

BidBert + BiLSTM + CRF 模型(这里的 BidBert 是在我们自己的招标数据集上使用 BERT 模型训练得到的)，其训练流程包括以下步骤：

S1、对纯文本、HTML 和 DOC 文档等进行数据预处理；

S2、将 S1 的结果输入 BidBert 模型得到每个句子中字或词的 Word Embedding；

S3、将 S2 的输出输入双向 BiLSTM 得到特征向量，再经过 CRF 对特征向量进行解码；

S4、通过 S2、S3 进行迭代训练，得到 BidBert + BiLSTM + CRF 模型。

3.4. 模型训练代码结构

```
bert_bilstm_crf_ner_pytorch
│   .gitignore
│   3.19.0
│   requirements.txt    # 环境包信息
│   README.md
└──torch_ner
    │   __init__.py
    ├──bert-base-chinese    #base Bert 模型
    │       config.json
    │       flax_model.msgpack
    │       pytorch_model.bin
    │       tf_model.h5
    │       tokenizer.json
    │       tokenizer_config.json
    │       vocab.txt
    ├──data # 训练、验证、测试数据集
    │       eval.txt
    │       test.txt
    │       train.txt
    │
    ├──output     # 模型保存路径
    │   ├──20220901174450
    │   │   │   config.json
    │   │   │   config.train
    │   │   │   label2id.pkl
    │   │   │   label_list.pkl
    │   │   │   ner_model.ckpt
    │   │   │   ner_model.pth
    │   │   │   pytorch_model.bin
    │   │   │   special_tokens_map.json
    │   │   │   tokenizer_config.json
```

```
│   │   │   token_labels_test.txt
│   │   │   training_config.bin
│   │   │   vocab.txt
│   │   └──eval
│   │           events.out.tfevents.1662025491.DESKTOP-STKHN7G
│   └──logs
│           ner_train.log   # 训练日志
├──source
│   │   config.py   # 配置文件，模型训练参数等在里面配置
│   │   conlleval.py   # 验证模型和测试模型类
│   │   logger.py   # 日志
│   │   models.py   # 主要模型，构建好的模型
│   │   ner_main.py   # 训练模型的入口
│   │   ner_main_bak.py
│   │   ner_predict.py   # 模型推理预测
│   │   ner_processor.py # 训练数据预处理
│   │   ner_test.py
│   │   pth2pt.py
│   │   test.py
│   │   utils.py
│   │   __init__.py
```

3.5. 模型应用

模型推理就是抽取标讯信息中的关键信息，是对纯文本、HTML 和 DOC 文档等中的关键信息进行抽取，其步骤包括：

S1、对纯文本、HTML 和 DOC 文档等进行数据预处理；

S2、将 S1 的结果输入到 BidBert + BiLSTM + CRF 模型中进行抽取，并得到结果。

通过对 1000 篇标讯信息的两种不同模型抽取核查，实验结果如表 1 所示，可以看出，BiLSTM-CRF 模型的各类命名实体识别准确率要大于 LSTM-CRF 模型，与单向的 LSTM-CR 模型相比，BiLSTM + CRF 模型准确率、召回率、F 值平均提高 1.13%、1.09%、1.22%。实验结果数据可以表明，将单向 LSTM 网络结构改进为双向 LSTM 网络结构，有助于对文本序列的上下文信息提取。

Table 1. Comparison of experimental results of different models of bids data
表 1. 标讯字段抽取不同模型实验结果对比

模型\实体	项目名称	项目编号	项目预算	项目 预算单位	报名 截止时间	投标 截止时间	业主	业主联系人
LSTM + CRF 模型准确率 P (%)	87.6	72.3	86.5	91.4	88.4	89.5	92.3	91.2
BiLSTM + CRF 模型准确率 P (%)	89.4	73.1	88.4	92.1	89.5	91.5	93.4	91.7
LSTM + CRF 模型召回率 R (%)	82.3	68.0	81.3	85.9	83.1	84.1	86.8	85.7

BiLSTM + CRF 模型召回率 R (%)	84.2	68.7	83.1	86.6	84.1	86.0	87.8	86.2
LSTM + CRF 模型 F 值(%)	85.8	70.1	83.9	88.7	85.7	86.8	89.5	88.5
BiLSTM + CRF 模型 F 值(%)	87.1	70.9	85.7	89.3	86.8	88.8	90.6	88.9
模型\实体	业主电话	代理机构	代理联系人	代理电话	中标机构	中标金额	中标金额单位	中标候选人
LSTM + CRF 模型准确率 P (%)	77.6	90.1	87.5	86.2	88.2	93.4	95.2	86.5
BiLSTM + CRF 模型准确率 P (%)	79.5	92.1	89.2	87.2	90.1	94.5	95.6	87.8
LSTM + CRF 模型召回率 R (%)	72.9	84.7	82.3	81.0	82.9	87.8	89.5	81.3
BiLSTM + CRF 模型召回率 R (%)	74.7	86.6	83.8	82.0	84.7	88.8	89.9	82.5
LSTM + CRF 模型 F 值(%)	75.3	87.4	84.9	83.6	85.6	90.6	84.1	76.4
BiLSTM + CRF 模型 F 值(%)	77.1	89.3	86.5	84.6	87.4	91.7	84.5	77.6

4. 结束语

BiLSTM + CRF 模型泛化能力强，缺点是需要大量的标注样本，采用迁移学习的思想，利用 Bert 在先验知识的基础上进行模型训练，BERT Embedding 与 BiLSTM CRF 分离[7]，可实现 GPU 显存的高效利用。

全国每天发布的招投标信息超过二十万条，绝大部分都是各种格式的复杂非结构化信息，如何识别及结构化，对于挖掘数据的隐藏价值具有非常重要的作用，本文尝试通过大量的文本标注，借助机器学习算法，不断优化模型，取得了较好的提取效果，对实现营销决策、数据统计及分析等各类大数据应用奠定了良好基础。

基金项目

贵州省科技计划项目(课题)黔科中引地[2021] 4016; 基于 BOES 开放引擎的数据分析关键技术在招投标领域的创新应用。

参考文献

[1] Zhang, Y. and Yang, J. (2018) Chinese NER Using Lattice LSTM. *Proceedings of the* 56*th Annual Meeting of the Association for Computational Linguistics*, Melbourne, July 2018, 1554-1564. https://doi.org/10.18653/v1/P18-1144

[2] Chen, D., Li, Z., Li, Z., *et al*. (2019) Semi-Supervised Entity Recognition of Chinese Government Document. *Proceedings of the* 2*nd International Conference on Artificial Intelligence and Pattern Recognition*, Beijing, 16-18 August 2019, 145-149. https://doi.org/10.1145/3357254.3357288

[3] 王子牛, 姜猛, 高建瓴, 陈娅先. 基于 BERT 的中文命名实体识别方法[J]. 计算机科学, 2019, 46(z2): 138-142.

[4] 赵畅, 李慧颖. 面向知识库问答的实体链接方法[J]. 中文信息学报, 2019, 33(11): 125-133.

[5] Cai, X., Dong, S. and Hu, J. (2019) A Deep Learning Model Incorporating Part of Speech and Self-Matching Attention for Named Entity Recognition of Chinese Electronic Medical Records. *BMC Medical Informatics and Decision Mak-*

ing, **19**, Article No. 65. https://doi.org/10.1186/s12911-019-0762-7

[6] Cheng, J., Pan, C., Dang, J., *et al*. (2020) Entity Linking for Chinese Short Texts Based on BERT and Entity Name Embeddings.

[7] Kareem, D., Ahmed, A., Hamdy, M. and Mohamed, E. (2020) Arabic Diacritic Recovery Using a Feature-Rich biLSTM Model. *ACM Transactions on Asian and Low-Resource Language Information Processing*, **20**, 1-18. https://doi.org/10.1145/3434235