中国科学院软件研究所学术年会’2021
暨计算机科学国家重点实验室开放周

学术论文

充分利用异构数据：一种解耦的两阶段训练的命名实体识别模型

Toward Fully Exploiting Heterogeneous Corpus: A Decoupled Named Entity Recognition Model with Two-stage Training

Yun Hu, Yeshuang Zhu, Jinchao Zhang, Changwen Zheng, Jie Zhou.

Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 1641–1652

Yun Hu (e-mail：huyun2016@iscas.ac.cn, tel：18811346717)

概述

- 通常的命名实体识别模型使用人工标注的数据，然而，数据标注需要耗费大量的时间和金钱，这限制了标注数据的规模，形成了命名实体识别模型的性能瓶颈。
- 在现实生活中，我们能够采用自动化的方法收集到大规模的实体词典和远程监督数据。但是，实体词典缺少有意义的上下文信息，远程监督数据包含大量的噪声，所以直接使用实体词典数据和远程监督数据会给命名实体识别模型带来不确定的因素。

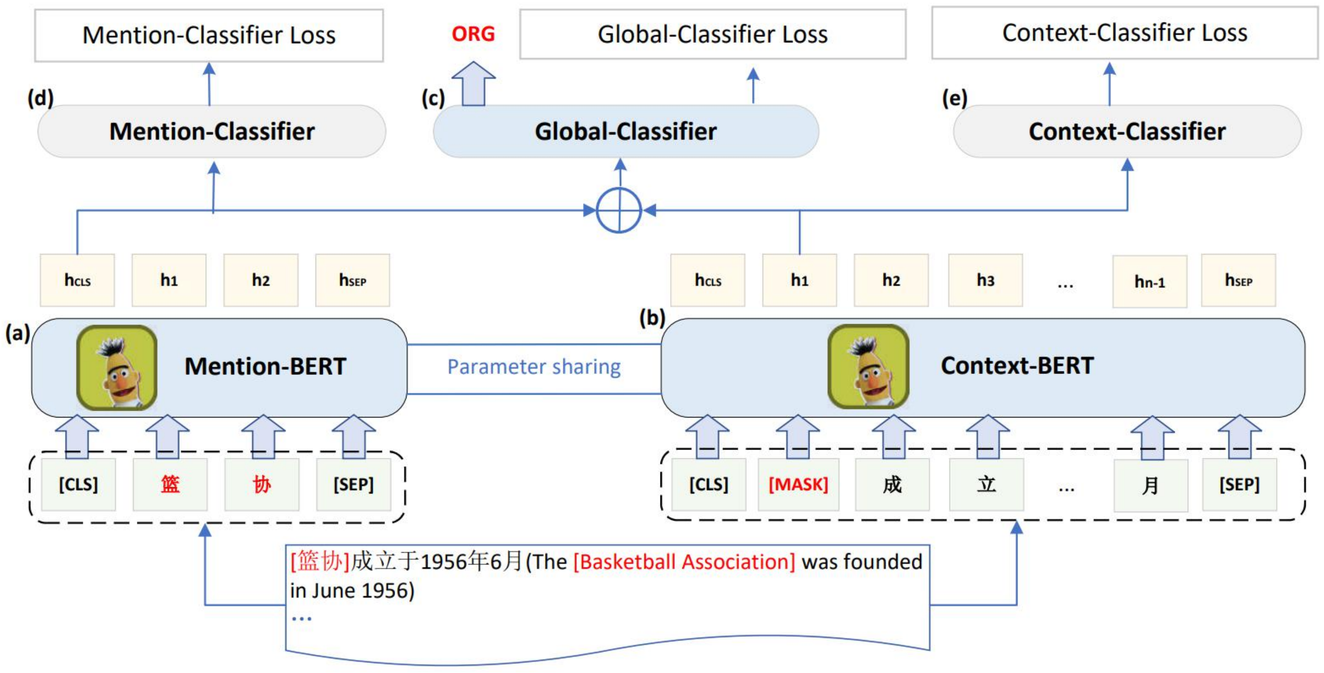
目标

- 更好的利用各种异构数据来提升命名实体识别的效果

思路

- 我们提出了一个基于BERT的二阶段解耦模型来更好的利用这些异构数据（实体词典，远程监督数据和人工标注的数据）。
- 在预训练阶段，我们设计了Mention-BERT和Context-BERT来分别学习上下文无关的实体词典和带噪声的远程监督数据。在微调阶段，通过对候选实体的预测，来对Mention-BERT和Context-BERT使用人工标注的数据进行统一训练。

模型架构



- 模型包括一个Mention-BERT，一个Context-BERT和一个Global-Classifier.
- 对于输入句子，“篮协成立于1956年6月”，首先会被转化为一个〈MENTION, CONTEXT〉对: 〈“篮协”，“[MASK]成立于1956年6月”>，之后“篮协”会作为Mention-BERT的输入，“[MASK]成立于1956年6月”会作为Context-BERT的输入，最后Mention-BERT和Context-BERT的输出会被连接，并且输入到Global-Classifier当中来得到最后的输出 (ORG).
- Mention-BERT希望能够捕获实体内部的结构信息，模型架构和传统的BERT相同，是Transformer的encoder部分

$$h_m = \text{Mention-BERT}(m)$$

- Context-BERT希望能够捕获实体的上下文信息，模型架构和传统的BERT相同，是Transformer的encoder部分

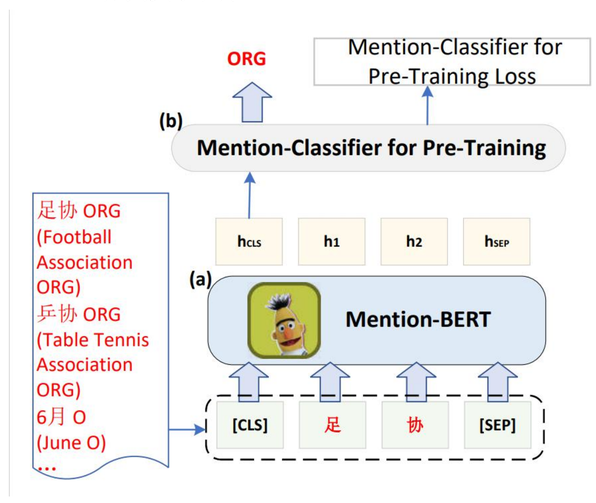
$$h_c = \text{Context-BERT}(c)$$

- Global-Classifier希望能够综合实体的内部结构信息和上下文信息，模型架构是一个全连接的网络

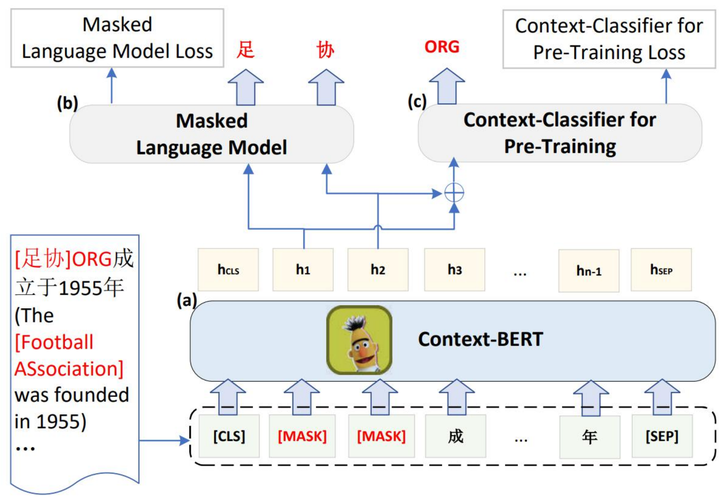
$$y_g = softmax(W_g \cdot [h_m : h_c] + b_g)$$

双阶段训练

- Mention-BERT的预训练



- Context-BERT的预训练



- 微调

$$L = L_g + \alpha L_m + \beta L_c$$

结果

- 我们在三个数据集上进行实验，使用P，R，F衡量效果

OntoNotes	P	R	F
BiLSTM-CRF (Lample et al., 2016)	68.79	60.35	64.30
Latice-LSTM (Zhang and Yang, 2018)	76.35	71.56	73.88
BERT-NER (Devlin et al., 2019)	78.01	80.35	79.16
Incomplete-NER (Jie et al., 2019)	79.18	81.24	80.20
MRC (Li et al., 2020b)	82.98	81.25	82.11
SoftLexicon (Ma et al., 2020)	83.41	82.21	82.81
FLAT (Li et al., 2020a)	-	-	81.82
CoFEE (Xue et al., 2020)	82.50	82.78	82.64
Decoupled model	83.79	83.06	83.43
+ Mention Pre-train	84.34	83.54	83.93
+ Context Pre-train	84.28	83.51	83.89
+ Mention and Context Pre-train	**84.92**	**83.72**	**84.32**

结论

- 我们提出一个解耦的双阶段训练的命名实体识别模型，能够更好的利用实体词典，远程监督数据和人工标记数据，在三个中文命名实体识别数据集的实验上，我们的方法都取得了最佳结果，并显著超越其他基线方法。