基于开源情报的某领域知识图谱构建

医疗领域中文命名实体抽取

答辩人：汪洪钧　指导老师：何亮

清华大学电子工程系

2021 年 1 月 6 日

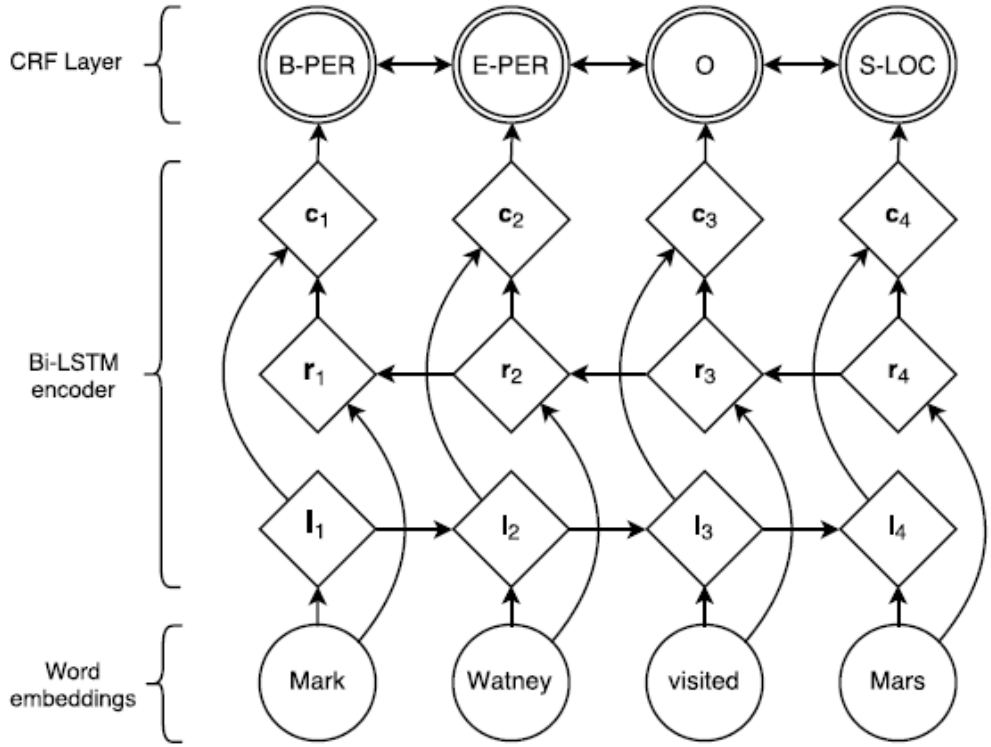
目录

命名实体识别方法

- 命名实体识别是信息抽取领域中的一项重要任务，也是许多下游智能应用的重要先决条件，例如决策系统和知识图谱的构造。
- 常见的命名实体识别任务：序列标注
- 命名实体识别方法：LSTM-CRF, BERT-LSTM-CRF 等
  - LSTM-CRF [1] 通过双向 LSTM 建模上下文关系，通过条件随机场建模标签之间的关系
  - BERT [2] 通过巨量语料库预训练的语言模型，可以针对各种下游任务进行微调，也可将输出接其他层进行训练

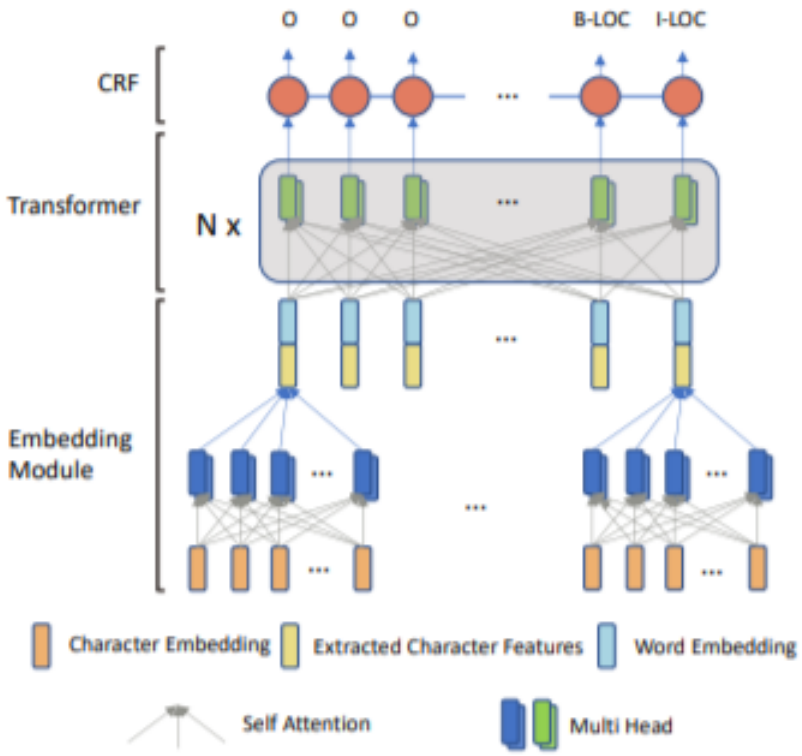
LSTM-CRF

- 使用 Bi-LSTM 学习上下文知识
- 使用 char-embedding 结合 world-embedding 作为 Bi-LSTM 的输入
- 使用 CRF 层处理 Bi-LSTM 的输出，得到结果



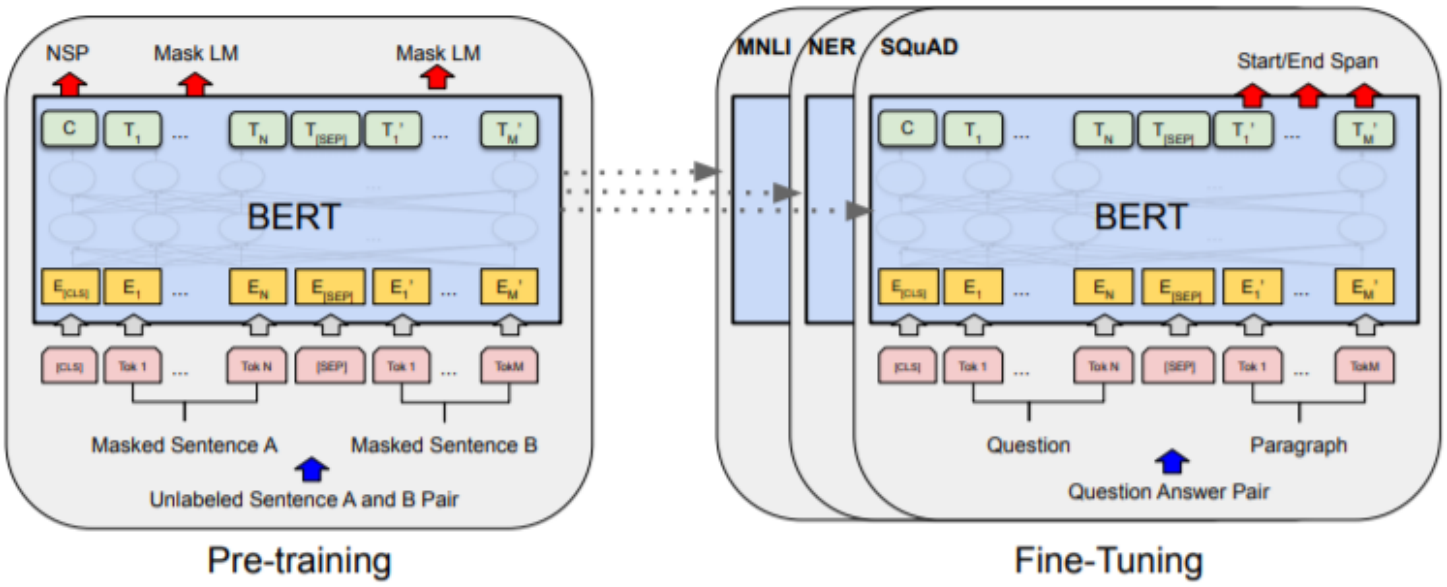
TENER

- Transformer 在 NER 任务上表现不佳：缺少方向信息
- TENER [3] 使用基于相对位置 (有正负) 的 Attention 机制

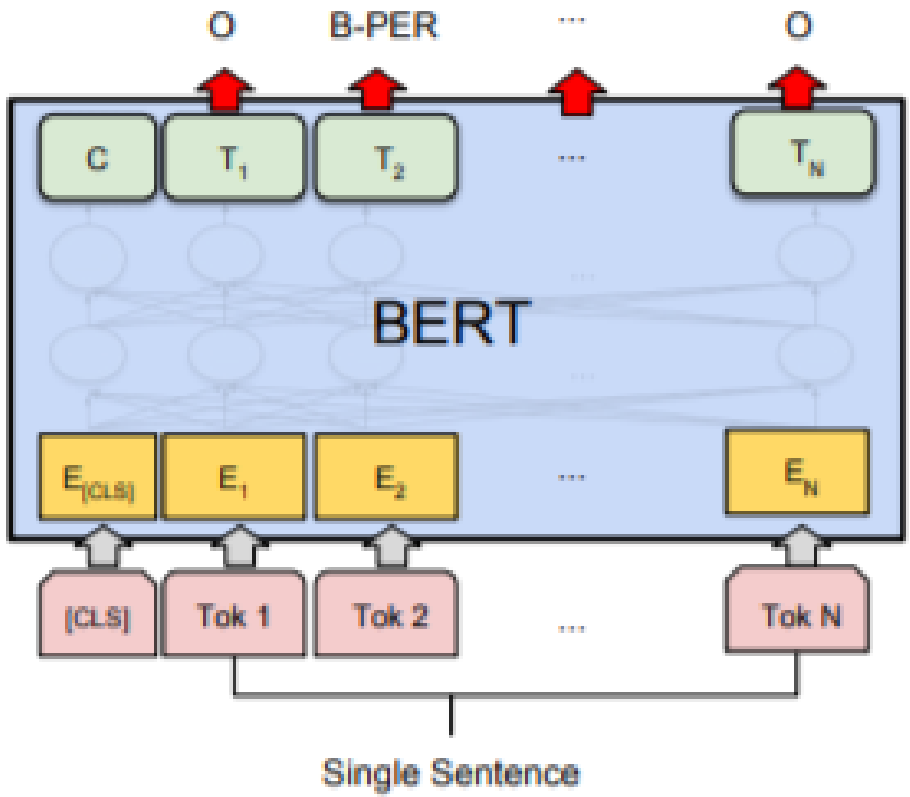


BERT

- 巨型的模型 (Transformer 24 layers, 1024 dim, 16heads) 与巨量的语料
- 训练任务一：随机掩蔽序列中的 tokens，进行预测
- 训练任务二：输入两个句子进行是否为连续句子的判断
- 在多个下游任务中，包括命名实体识别取得优秀成果



BERT


O
B-PER
...
O
C
T1
T2
...
TN
BERT
E(CLS)
E1
E2
...
EN
[CLS]
Tok 1
Tok 2
...
Tok N
Single Sentence

课题内容

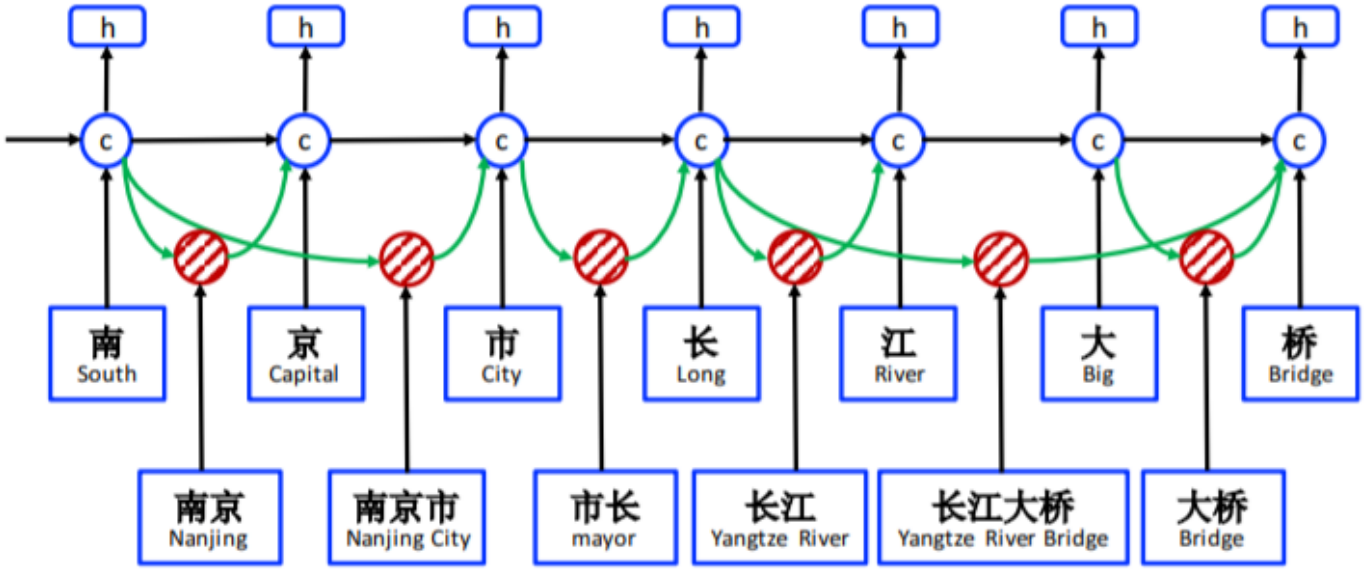
1. 医疗领域中文命名实体识别
2. 数据集
3. 系统实现

中文命名实体识别的难点

- 中文命名实体识别的难点在于" 词" 的划分
  - 中文序列的独特性在于其输入的单位是" 字"，难以实现字词结合
  - 例如：王小美今天坐高铁来到北京南站。
  - 经过中间步骤" 分词" 的效果不如仅仅以" 字" 为单位输入进行 LSTM-CRF
- Lattice LSTM [4] 是词信息引入的结构的开山之作

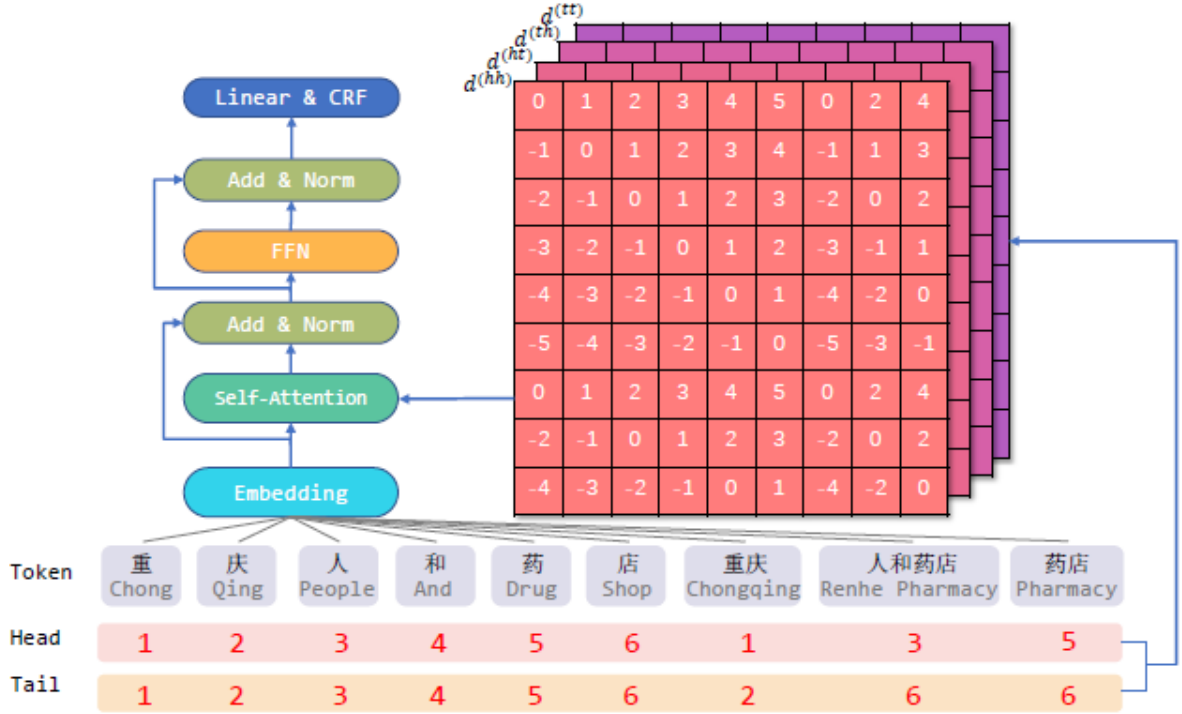
Lattice LSTM

- 在计算第 k 个单元时，除了常规 LSTM 的输入外，还考虑以位置 k 结尾的词的信息
- 同样地，使用 Bi-LSTM 和 CRF
- 缺点：无法并行计算，难以移植到其他结构



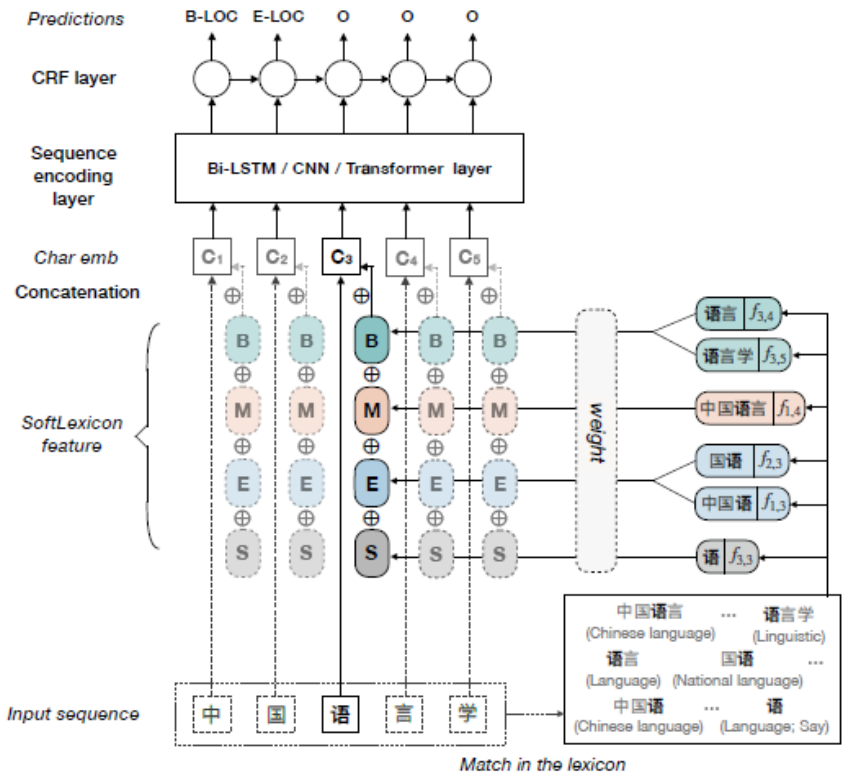
FLAT

- FLAT [5] 的想法来自于 Lattice 和 TENER 等
- 将字和词都作为 Transformer 的输入，同时将它们的首尾 (相对) 位置也作为输入，进行基于相对位置的 Attention
- 仅使用一层 encoder



Soft-Lexicon

- Soft-Lexicon [6] 同样是词信息的嵌入，使用的是词语的分类嵌入拼接



针对中文医疗领域

- 将尽量综合以上各个结构的优点进行系统融合
- 小领域的实体抽取难点在于语料的缺乏
  - 将考虑实体词典 (Gazetteers，含实体类型标签) 强化语料
- 会考虑采用一些新的结构
  - 更注重推理和全局信息的图结构
  - 使用新的预训练语言模型，如 LUKE [7],XLNet [8]

中文数据集

- 常见的中文数据集有以下四个

	Ontonotes	MSRA	Resume	Weibo
Train	15740	46675	3821	1350
Char_{avg}	36.92	45.87	32.15	54.37
Word_{avg}	17.59	22.38	24.99	21.49
Entity_{avg}	1.15	1.58	3.48	1.42

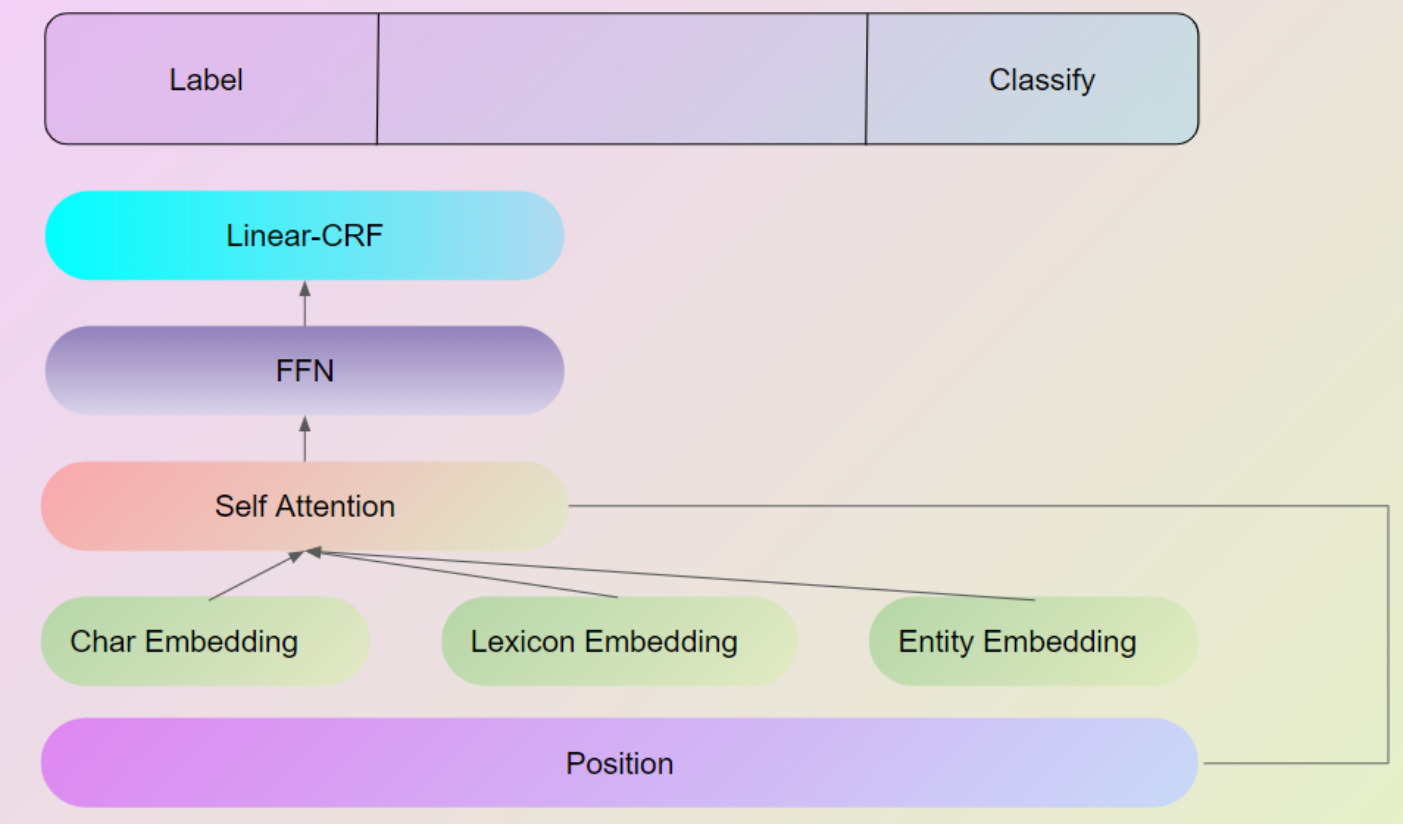
- 词信息来自 Lattice-LSTM [4] 提供

医疗领域数据集

- 医疗领域数据集
  - 具体的评测任务提供 (如:CHIP2020)
  - CCLUE：中文临床自然语言处理算法评估基准
- 实体词典可以来自该领域相关的平台 (比如医疗领域有医渡云)
- THUOCL 提供医学类词信息

系统实现

- 拟定采用实体词典强化实体识别，结构化信息强化实体分类的多任务结构



- 拟定采用 Lattice-LSTM [4] 作为 baseline

计划进度

- 2020.10-2021.12：文献调研
- 2021.01-2021.03：在数据集上跑通 Baseline，以及其他有较好表现的系统，然后搭建自行设计的系统框架，运行系统
- 2021.04-2021.05：验证结果

参考文献 I

[1] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer, "Neural architectures for named entity recognition," arXiv preprint arXiv:1603.01360, 2016.

[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," arXiv preprint arXiv:1810.04805, 2018.

[3] H. Yan, B. Deng, X. Li, and X. Qiu, "Tener: Adapting transformer encoder for name entity recognition," arXiv preprint arXiv:1911.04474, 2019.

[4] Y. Zhang and J. Yang, "Chinese ner using lattice lstm," arXiv preprint arXiv:1805.02023, 2018.

[5] X. Li, H. Yan, X. Qiu, and X. Huang, "Flat: Chinese ner using flat-lattice transformer," arXiv preprint arXiv:2004.11795, 2020.

[6] M. Peng, R. Ma, Q. Zhang, and X. Huang, "Simplify the usage of lexicon in chinese ner," arXiv preprint arXiv:1908.05969, 2019.

[7] I. Yamada, A. Asai, H. Shindo, H. Takeda, and Y. Matsumoto, "Luke: deep contextualized entity representations with entity-aware self-attention," arXiv preprint arXiv:2010.01057, 2020.

[8] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le, "Xlnet: Generalized autoregressive pretraining for language understanding," in Advances in neural information processing systems, pp. 5753–5763, 2019.

谢谢！