周报 2023/02/03

文献阅读

%使用标准文献引用格式，列出几篇本周精度的文献%

例：

[1] Zhang Y, Yang J. Chinese NER using lattice LSTM[J]. arXiv preprint arXiv:1805.02023, 2018.

[2] Li X, Yan H, Qiu X, et al. FLAT: Chinese NER using flat-lattice transformer[J]. arXiv preprint arXiv:2004.11795, 2020.

[3] Gu, Y., Qu, X., Wang, Z., Zheng, Y., Huai, B., & Yuan, N. J. (2022). Delving deep into regularity: a simple but effective method for Chinese named entity recognition. arXiv preprint arXiv:2204.05544.

研究进展/文献感悟/科研想法

文献 1：Lattice LSTM 这篇论文发表在 ACL 2018 的会议上，提出 Lattice LSTM 模型。文章中使用了一个大规模的自动获取的词典与句子进行匹配，得到词-字符格，由此每个字符的向量表示包含字符本身的信息和包含这个字符的所有词的信息。与 character-based 的方法相比，这个模型显式的利用了词的信息。与 word-based 的方法相比，不会出现分词错误。这篇文章是在 NER 任务中引入词典的方法的开山之作。

文献 2：第二篇文章是第一篇文章的改进版本，提出了基于 Transformer 来解决文献 1 中的词汇损失问题（每个字符只能获取以它为结尾的词汇信息），并且使用相对位置编码来使 Transformer 适应 NER。模型具体为每个字符和每个潜在的 word 使用 head 和 tail 两个索引去表示 token 在输入序列中的绝对位置，head 表示开始索引，tail 表示结束索引，对于每个字符，head 和 tail 是一样的，而每个 word 的 head 和 tail 是不一样的。采用了 Transformer 结构，相比于 LSTM 最大的优势是可以加速计算。

文献 3：对于 NER 问题有两种建模方式，sequence labeling 的方式和 span_based 的方式，而 sequence labeling 的方式不能很好地识别嵌套的实体，也就是一个实体存在于两个实体中时仅能识别出一个实体。这篇文章就是基于 span_based 的方法。

科研想法：对于流调报告中的实体识别考虑引入地名词典，采用 lexicon_enhanced 方法

下周计划

寻找是否存在地名词典，验证 lexicon_enhanced 方法的可实现性，撰写论文的 literature review 部分

存在问题

实验部分代码能力还需加强，阅读代码速度较慢