$$df(w,c)=\frac{\text{词w在类别c中的样例数}}{\text{类别c的总样例数}}$$

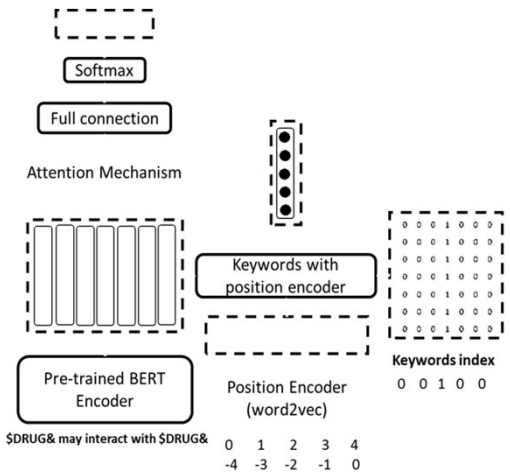
结合类别关键词与注意力机制的药物关系抽取

蔡晓玲　Ika Novita Dewi　董守斌
（华南理工大学 计算机科学与工程学院，广东 广州 510000）

论文摘要

在深度模型中有效利用类别关键词，挖掘类别关键词与药物实体的关系，及关键词与句子中其他词的关联关系，可望增加样例区分度和提高模型分类效果。本文基于卡方检验和文档频率获取每个类别的关键词，在预训练BERT模型中加入关键词与药物对的位置编码来增加样例的差异性，并通过注意力机制学习关键词与句子中其他词的分布信息。同时，针对药物关系抽取任务中负样例较多的问题，提出了基于规则和模式的负样例过滤方法，有效降低正负样本比例。本文方法在公开DDI数据集上取得了82.41%的F1值，是目前在该数据集上的SOTA结果

系统模型



算法原理

- 通过卡方检验以及文档频率筛选类别关键词
  - 本文基于词与类别相互独立的假设，计算实际观测值与理论推断值之间的偏离程度。偏离程度越大，说明假设不成立，词与类别相关。过滤文档频率小于0.05的词，取每个类别卡方值排前10的关键词，公式如下

$$\chi^2(w,c)=\frac{N\times(AD-CB)^2}{(A+C)\times(B+D)\times(A+B)\times(C+D)}$$

$$df(w,c)=\frac{\text{词w在类别c中的样例数}}{\text{类别c的总样例数}}$$

- **BERT预训练词向量**
  - 通过BioBERT获得句子的词向量
- **关键词向量与位置编码**
  - 通过keywords Index获得句中的关键词向量，并进行加和平均得到一个平均词向量v
  - 对句中实体对与句中各个词进行相对位置编码
- **基于关键词的注意力机制**
  - 利用attention机制计算关键词与句子中其他词的关联关系，给不同词分配不同的权重，来获得关键词与其他词的共现信息，然后进行分类

实验仿真

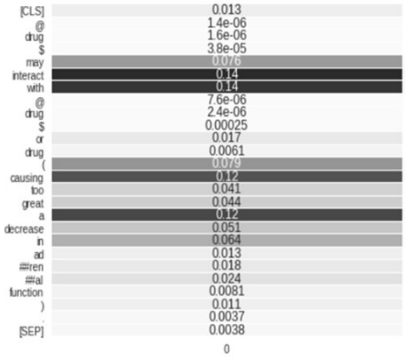
- **在DDI数据集上的模型性能结果**

模型	P(%)	R(%)	F1(%)
CNN	75.7	64.6	69.75
MCCNN	75.99	65.25	70.21
DCNN	78.24	64.66	70.81
DLSTM	72.5	71.5	72.0
PM-LSTM	75.80	70.38	72.99
Att-LSTM	78.4	76.2	77.3
BERT	80.65	78.35	79.48
R-BERT	81.13	79.97	80.55
BERT+Gaussian	**83.28**	79.88	81.54
My work	82.34	**82.48**	**82.41**

- **在DDI数据集上的消融实验**

方法	P(%)	R(%)	F1(%)
基准模型	80.67	78.17	79.39
+负样例过滤	80.05	81.35	80.70
+类别关键词	81.83	81.05	81.44
+位置编码	82.34	82.48	82.41

- **Attention热力图分析**
  - Int类型的样例有两个关键词，分别是“int”类的“interact”和“effect”类的“decrease”。模型更关注的interact这个关键词



论文结论

基于卡方检验和文档频率来获取类别关键词，在模型中增加对关键词的位置编码，并通过注意力机制学习关键词与其他词的分布信息。在实验中，将本文工作与其他经典模型进行对比，证明了模型的有效性，在药物关系抽取任务上取得了SOTA结果。本文还测试了各部分工作对模型性能的提升，并且通过attention热力图直观展示了模型效果。

CCNL 广东省计算机网络重点实验室 Communication & Computer Network Lab of GD