医学领域命名实体识别

1. 背景综述

众所周知，患者的电子病历贯穿医疗活动的始终，是医疗信息系统的核心数据。电子病历（英文：Electronic Medical Record，EMR）是指医务人员在医疗活动过程中，使用医疗机构信息系统生成的文字、符号、图表、图形、数据、影像等数字化信息，并能实现存储、管理、传输和重现的医疗记录，是由医务人员撰写的面向患者个体描述医疗活动的记录。我国医疗机构数量庞大，患者的就医需求也与日俱增，门诊病历和住院病历急剧增长。电子病历由医务专业人员撰写，不仅仅是具有法律效力的医疗活动证据，而且包含大量的专业医疗知识。通过分析电子病历能挖掘出这些与患者密切相关的医疗知识。比如，某患者电子病历中，“头 CT 检查显示腔隙性脑梗死”。在这句话中，“头 CT”是检查手段，“腔隙性脑梗死”是疾病，这二者在电子病历信息抽取研究中被称为命名实体或概念，这两个实体间的关系是“头 CT”证实了“腔隙性脑梗死”的发生，或者说“腔隙性脑梗死”可以通过“头 CT”这种检查手段得到确认。从电子病历里自动挖掘这些知识就是要自动识别电子病历文本中与患者健康密切相关的各类命名实体以及实体间的关系。并且信息抽取命名实体为后续的各种文本挖掘任务提供标准和便利。

电子病历的命名实体识别虽然前景诱人，但是做起来也会面临着相当大的难度，需要大量的医学术语及电子病历的标注数据、高性能的深度学习服务器，这两个必须点都是我们没有的，我们只能做尝试性的工作。

2. 技术路线

2.1. 技术架构

用 IDCNN 和 BI-LSTM 做端到端的医学领域中文实体识别：

```
embedding + BI-LSTM/IDCNN + crf + softmax
```

2.2. 语料获取

互联网爬虫抓取数据，人工标注数据。

医疗活动主要涉及四类重要信息：症状、疾病、检查和治疗，涉及的具体描述如下：

- 疾病：泛指导致患者处于非健康状态的原因，比如：诊断、病史。
- 症状：泛指疾病导致的不适和显示表达的检验检查结果，分为：自诉症状和体征（异常检验检查结果）。
- 检查:泛指为了得到更多的由疾病导致的异常表现以支持诊断而采取的检查程序、检查项目等。
- 治疗：泛指为了治愈疾病、缓解或改善症状而给予患者的药物、手术和措施等。

2.3. 数据标注

选用 IOBES 序列标注方法，标注标签如下：

- B，即 Begin，表示开始
- I，即 Intermediate，表示中间
- E，即 End，表示结尾
- S，即 Single，表示单个字符
- O，即 Other，表示其他，用于标记无关字符

2.4. 模型训练

2.4.1. 系统要求

- Python(>=3.5)
- TensorFlow(>=r1.0)
- Jieba(>=0.37)

2.4.2. 模型示意图

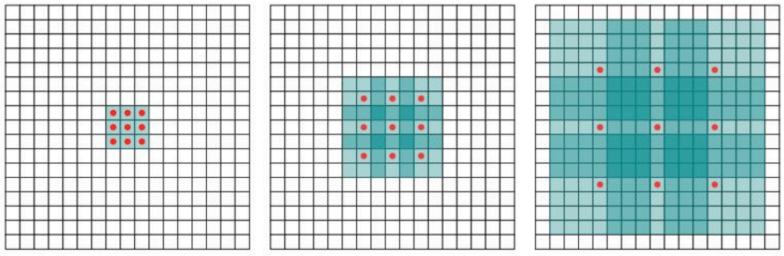

Figure 1: Systematic dilation supports exponential expansion of the receptive field without loss of resolution or coverage. (a) F_1 is produced from F_0 by a 1-dilated convolution; each element in F_1 has a receptive field of 3×3. (b) F_2 is produced from F_1 by a 2-dilated convolution; each element in F_2 has a receptive field of 7×7. (c) F_3 is produced from F_2 by a 4-dilated convolution; each element in F_3 has a receptive field of 15×15. The number of parameters associated with each layer is identical. The receptive field grows exponentially while the number of parameters grows linearly.

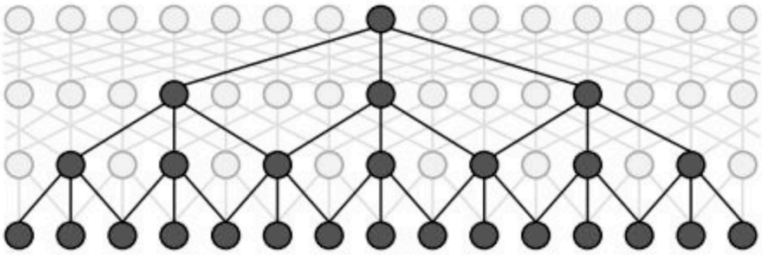

Figure 1: A dilated CNN block with maximum dilation width 4 and filter width 3. Neurons contributing to a single highlighted neuron in the last layer are also highlighted.