# 病历实体识别（CCK2017任务2）失败总结

July 12, 2017

## 1 任务简述

这是一个识别实体和实体分类的任务。我们拿到的数据见Table 1，任务就是将原始语料中的实体位置识别出来，以及对实体进行分类。训练语料一共1198条，测试语料398条。训练语料上有正确的实体位置，测试语料未经标注。

| 原始语料 | 主因上腹部、腰部疼痛1天入院。 | |
|---|---|---|
| 实体 | 位置 | 种类 |
| 上腹部 | 2，4 | 身体部位 |
| 腰部 | 6，7 | 身体部位 |
| 疼痛 | 8，9 | 症状和体征 |

Table 1: 原始训练语料。

## 2 基于词的标注方式

这个任务有两个子任务，实体识别和实体分类，我们用一个模型以序列标注的方式来一起做这两个子任务。

我们最开始用了基于词的做法，这种方法在本次任务中就会出现一个很严重的问题。首先是用分词器将原始语料分词，但是分词器可能对实体的边界的分词处理并不好，比如可能将Table 1 中的原始语料分割如下分词结果:

主　因上　**腹部**　、　**腰部疼**　**痛**　1　天　入院　。

这样的分词结果没有办法对粗体的实体进行正确的标注。我们虽然能在训练集上使用了补救的方法，可以先用实体分割原始语料，得到这样的结果:

主因　**上腹部**　、　**腰部**　**疼痛**　1天入院。

然后对分割后的结果逐一进行分词，最后对其结果进行序列标注，标注例子见Figure 1，但是在测试集，我们没有办法先用实体进行切分，因为我们不知

道测试集上的实体位置。这也就导致了测试集和训练集分词风格有所不同，如此一来有部分实体在测试集上不可能有正确的标注。

在这种基于词的方式下，两次有效的提交结果性能见Table 2。我们在基本模型（神经网络特征+稀疏特征+CRF）上再加入字特征，可以看到字的特征对基本模型性能提升没有太多效果。甚至在严格的评价方式下性能有所下降，出现了过拟合的现象。可见这个任务在这种基于词的标注方式下，几乎没有机会用字特征提升模型性能。

| 主 | 因 | 上 | 腹部 | 、 | 腰部 | 疼痛 | 1 | 天 | 入 | 院 | 。 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| o | o | b-stbw | e-stbw | o | s-stbw | s-zztz | o | o | o | o | o |

Figure 1: 基于词的标注结果

| 模型 | Relaxed F-score | Strict F-score |
|---|---|---|
| 基本模型 | 85.2 | 76.6 |
| 基本模型加字特征 | 86.8 | 76.4 |

Table 2: 组织方对我们自动标注的测试集总体性能评价结果。

可能有人会对我们使用的基本模型加入字特征是否有效提出质疑，我们用分词完全正确的训练集进行了如下实验。我们把1198句训练语料分割998句训练集，100句开发集，100句测试集。测试结果如Table 3所示。

| 模型 | P | R | Strict F-score |
|---|---|---|---|
| 基本模型 | 87.7 | 88.8 | 88.3 |
| 基本模型加字特征 | 88.7 | 90.8 | 89.7 |

Table 3: 分词完全正确的情况下测试集上的性能评价结果。

# 3 基于字的标注方式

基于字的序列标注方式可以避开上述的实体的边界问题，关于字的序列标注见Figure 2。在这种标注形式下，组织方对我们自动标注的测试集给出评价结果，见Table 4。可以看到我们的模型性能有明显的提升。


Figure 2: 基于字的标注结果

| 模型 | Relaxed F-score | Strict F-score |
|---|---|---|
| 基于词模型 | 85.2 | 76.6 |
| 基于字模型 | 91.9 | 84.8 |

Table 4: 组织方对我们自动标注的测试集总体性能评价结果。