NER标注规则

标注规则中，所有的O都代表Other，即该标签对应的词不属于实体的一部分。

**IOB1:**

使用连续的I来标注实体中的词，使用B作为边界来区分无缝连接的多个完整的实体

例：

source: 感冒 很好 治疗

label: I-DIS O O

DIS entity: 感冒

source: 风湿性 关节炎 很难 痊愈

label: I-DIS I-DIS O O

DIS entity: 风湿性关节炎

source: 我 得了 感 冒 发 烧 肠 胃 炎

label: O O I-DIS I-DIS B-DIS I-DIS B-DIS I-DIS I-DIS

DIS entity: 感冒、发烧、肠胃炎

注：不同实体类型（即标签的后缀不同）不需要使用B来区分

source: 我 得了 感 冒 打 喷嚏

label: O O I-DIS I-DIS I-SYM I-SYM

DIS entity: 感冒

SYM entity: 打喷嚏

**IOB2:**

使用一个B作为实体的起始标志，连续的I作为实体中后续的词的标志

例：

source: 感冒 很好 治疗

label: B-DIS O O

DIS entity: 感冒

source: 风湿性 关节炎 很难 痊愈

label: B-DIS I-DIS O O

DIS entity: 风湿性关节炎

source: 我 得了 感 冒 发 烧 肠 胃 炎

label: O O B-DIS I-DIS B-DIS I-DIS B-DIS I-DIS I-DIS

DIS entity: 感冒、发烧、肠胃炎

source: 我 得了 感 冒 打 喷嚏

label: O O B-DIS I-DIS B-SYM I-SYM

DIS entity: 感冒

SYM entity: 打喷嚏

**IOBES:**

使用一个B作为实体的起始标志，一个E作为实体的终结标志，B和E之间若还有其他属于该实体的词，都使用I来标记；单独一个词成实体的使用S来标记。

例：

source: 感冒 很好 治疗

label: S-DIS O O

DIS entity: 感冒

source: 风湿性 关节炎 很难 痊愈

label: B-DIS E-DIS O O

DIS entity: 风湿性关节炎

source: 我 得了 感 冒 发 烧 肠 胃 炎

label: O O B-DIS E-DIS B-DIS E-DIS B-DIS I-DIS E-DIS

DIS entity: 感冒、发烧、肠胃炎

source: 我 得了 感 冒 打 喷嚏

label: O O B-DIS E-DIS B-SYM E-SYM

DIS entity: 感冒

SYM entity: 打喷嚏

Bi-LSTM-CRF-pos代码中，有特定的功能函数对以上标注规范进行转换，所以理论上来讲采用哪种标注规范都可以进行互转，你们看哪种标注规范方便人工来标记就采用哪种就可以。

utils.py文件中，iob(tags)函数的功能是检查数据的标注规则是否属于IOB1或IOB2；并且若数据标注规则使用了IOB1，该函数会自动将其转换为IOB2的标注规则再进行后续处理。iob\_iobes(tags)函数的功能是将IOBES的标注规范转换为IOB2，iobes\_iob(tags)函数的功能是将IOB2的标注规范转换为IOBES。