命名实体抽取说明

1．需要抽取的什么

命名实体抽取，不仅要求能从文本中识别出具体的命名实体，而且还需要识别命名实体之间存在的关系。这里主要要求能够抽取三种命名实体和三种关系，对于命名实体而言，包括人名、地名以及机构名称；而对于关系而言，则包括人和机构之间的关系（有两种，即 founder 和 employee 关系）以及机构和地方之间的关系（当然，需要相关的描述，如 located in, founded in, registered in 等）。

2．命名实体及关系表示

在训练语料中，会分别用 nr、ns 和 nt 来表示人名、地名及机构名称。如在句子“**{湖南/ns}{郴州市/ns}{刘洪杰/nr}一日，我到民政部门办点事，刚一坐下，就进来一位来访的乡下老人。**”中，{湖南/ns}、{郴州市/ns}表示湖南、郴州市是地名，{刘洪杰/nr}则表示刘洪杰是个人名。而对于关系的表示，并没有直接在训练数据中标出，不过可以从“中文”和“英文”两个目录下的 xml 文档中看到关系的具体表示。下面我将会具体介绍如何通过输出文档的形式来表示抽取出命名实体和关系。

3．训练数据

人工标注好的训练数据放在文件夹“中文”和“英文”的文件 TrainingData.txt 中，**该文件中的每一行是一篇文档**，因此很容易看到中文训练语料共有 1000 篇，而英文训练语料则只有不足 30 篇（对此，实在抱歉）。而且训练数据中的相关命名实体都已经按照上面 2 的说明全部标出，同学们可以利用已有的训练数据进行模型学习。

4．输出格式

为了规范输出格式，以便于对同学们的结果进行评测，因此这里很有必要对测试结果的输出格式进行具体介绍。

在目录“中文”和“英文”下都有两个文件，即 SampleTestData.txt 和 SampleTestResult.xml，其中 SampleTestData.txt 是一个包含了未标注数据的样例测试文件，其中共有 5 篇文章，而 SampleTestResult.xml 文件则是这个测试文件的、类似于标准答案的输出。换句话说，到时候我在检测你们的系统或模型时，会给你们一个近似 SampleTestData.txt 的文件，该文件中包含了测试数据，而你们则需要输出一个同 SampleTestResult.xml 结构一样的文件，该文件包含了你们的测试结果。

SampleTestData.txt 文件很简单，仅仅包含了数篇没有进行标注的测试文章，再次强调，每一行表示一篇文章。而 SampleTestResult.xml 文件的结构则似乎需要徐徐道来。不过我想，对 XML 熟悉的同学，对这个文件的结构该是很容易理解的。

我们知道，每一个XML文档必须具有且仅有一个根元素，而且每个元素必须具有tag name，没理由我们不将根元素的tag name定为EntityExtraction。想到我们既要抽取实体，也要抽取实体之间的关系，因此，我为根元素添加了两个子元素，即Entities和Relations。接下来就简单说明下这两个元素吧。

a) Entities元素

当你展开Entities元素时，会发现它有一个count属性，该属性记录着从测试文件中提取出的命名实体数量。同时，Entities元素下有count个子元素，这些子元素具体地描述了每个命名实体的具体信息。不妨从SampleTestResult.xml文件中任意摘一个过来瞅瞅其模样。

```
<Entity type="nr">
  <Text>毛泽东</Text>
  <Doc>1</Doc>
  <Start>536</Start>
</Entity>
```

注意到了，每个Entity元素有个type属性，即类别属性，可以为nr(人名)、ns（地名）以及nt(机构名称)，type只能包含一个枚举值。同时每个Entity元素还有三个子元素，Text元素表示这个Entity的名字，Doc表示它在哪篇文章中出现（**文章序号从1开始**，也可以认为是测试文件中的行号），Start元素表示它在文章中哪个地方出现（即从第几个字符开始，**字符序号从0开始**）。因此，Entity元素能够可靠地确认测试数据中的一个命名实体了。需要说明的是，如果同一个命名实体在一篇文章中出现多次的话，你必须构建多个这样的Entity元素，这些元素可能唯一不同的地方就在于Start元素的取值了。

b) Relations 元素

同 Entities 元素一样，Relations 元素也有一个 count 属性，表示从测试数据中抽取出的关系个数，这里让我们来看下“英文”目录下的文件 SampleTestResult.xml 中的 Relation 元素的具体格式：

```
<Relation type="nr-nt">
  <NR>Paul Allen, Bill Gates</NR>
  <NT>Microsoft</NT>
  <Doc>3</Doc>
  <Desc>founder</Desc>
 </Relation>
```

也同 Entity 元素一样，Relation 也有一个 type 属性，即表示关系的种类。属性 type 的值可以为“nr-nt”或者“nt-ns”，“nr-ns”表示的是人和机构之间的关系，“nt-ns”表示的是组织机构同地理位置间的关系。每个 Relation 元素都包含四个子元素，但 type 取值不同时，包含的元素类型类有不同，这里我们分别来讨论。当 type 为“nr-nt”时，Relation 元素需包含 NR、NT、Doc 和 Desc 元素，NR 表示了人名，NT 表示了机构名称，Doc 则表示这个关系在哪篇文章中出现，Desc 则对这种关系进行描述，我们只需抽取人和机构之间的 founder 和 employee 关系，因此 type 为“nr-nt”时，Desc 仅能取值 founder 或 employee。而当 type 为“nt-ns”时，Relation 元素则需包含 NT、NS、Doc 和 Desc 元素，NT 表示了人名，NS 表示了机构名称，Doc 和 Desc 含义未变，但当 type 为“nt-ns”时，Desc 值需要通过文章中语境确定，如可以为“registered in”（NT 在 NS 注册），“headquatered in”（NT 总部位于 NS）等。

我想通过对上面 XML 文档格式的介绍，大家都具体的输出格式应该很清楚了把。

5．文件编码格式

“中文”和“英文”目录下的文件，出了 XML 文档采用 UTF-8 编码外，其余文件均采用的是 GB2312 格式。到时候提供的测试数据，同样也是 GB2312 格式的，不过你们的 XML 输出文档最好采用 UTF-8 编码。