**Species识别实验报告**

刘景琛

2010年1月18日

这次实验主要测试基于字典匹配的物种识别方法的性能。物种识别是Gene Normalization的一个重要步骤，通过识别出和基因蛋白质同现的物种，可以利用物种信息对含有歧义的基因蛋白名字消歧。虽然使用的方法是最简单的字典匹配方法，但是实验结果表明，由于文本中物种出现的形式比较简单单一，所以简单的字典匹配方法就已经可以取得很好的效果。

1. **实验数据**

实验使用的数据时”Disambiguating the Species of Biomedical Named Entities using Natural Language Parsers”一文中提供的数据，是BioCreative I和II的全部Gene Normalization任务的数据，共644篇摘要。他们的标注是将每篇摘要中的每个基因蛋白名字标出并且给出这个名字所在的物种的NCBI的id。通过这个标注，我们就可以知道每篇摘要中含有的物种。我们通过字典匹配，得到每篇摘要中的物种后，得到该篇摘要的物种集合，就可以和标注中得到的物种集合对比，得到precision、recall和F。

1. **实验方法**

使用的方法就是字典匹配方法，和蛋白质基因的字典匹配完全一样，也是对字典中每个词进行倒排文档式的索引，然后对每个句子中的每个token，得到含有它的字典记录集合，在对相邻token根据相同的字典记录进行合并，得到识别出的名字，最后在根据字典和简单规则进行后处理过滤。

所使用的字典是丁石林他们留下的物种字典，其格式为：

NCBI@9606:Homo sapiens~human~man~

前面是NCBI的物种id，后面就是一些名字和描述。

我自己又向里面加了一点儿内容，比如在老鼠那一条加了murine，在果蝇那一条加了Drosophila和fly，这些都是一些常见的写法，原来的字典上却没有。

在匹配中，将单词转化成单数形式比较重要，很多地方出现humans，以及mouse的复数mice。

1. **实验结果和分析**

根据上述实验方法和评测方法进行实验，得到的实验结果为：Precision 0.6335 Recall 0.8046 F 0.7089，但是这个结果并不是很准确的。

在precison方面。其实物种的名字是形式非常简单，就是human、mouse这样的简单单词或词组，只要用字典识别出来了基本上就不会错。得出的结果一共是2159个，我人眼看了一遍，只有27个我觉得是错了，而且这些错误大部分就是两种：把“spot”当成了一种鱼，把蛋白质“CAT”当成了猫。这样的错误哪怕用词表过滤也能解决。Precision只有0.6的原因是由于标注数据中只给出了有蛋白出现的物种，但摘要中有很多物种并没有蛋白出现在文中。比如有很多篇数据中只有human这一个物种，但实际上里面也提到了果蝇、老鼠，只是没有对应的蛋白。

在recall方面，虽然0.8对字典匹配也算是比较好的，但我还是发现很多False Negetive并不是我们的错误。我找出了一些我们没有识别出来的结果所在的原文，读了一下，也没发现标注中的物种，可能有些标注是根据蛋白质得出的物种，而不是原文中直接出现了物种。比如有些摘要完全没有任何与物种有关的词出现，但是在标注中蛋白质的物种都是人类。

这次物种识别的实验虽然使用的数据不是很合适，但是经过实验和分析还是能够看出，物种的识别比基因蛋白质识别简单很多。物种识别通过字典方法就能取得不错的结果，如果有比较好的训练数据，使用CRF等学习方法，效果会更好。目前物种识别这方面的主要问题还是没有合适的训练和测试数据。

之后我有尝试了增加从蛋白质名字中得到的物种，例如“hZIP2”这样的。蛋白质名字我直接从数据的标注结果中得到，因为我们主要是测试物种识别的性能，这样就去除了我们的蛋白质名字识别系统的结果与标注结果不同带来的影响。对于每个蛋白质，要求其第一个字母为小写h/r/y/m/d，并且第二个字符是大写的。再根据首写字母确定NCBI的物种ID，如表1。这样加入之后，结果变为Precision 0.6350 Recall 0.8072 F 0.7108，只有微弱的提升。

|  |  |  |
| --- | --- | --- |
| 字母 | 物种 | ID |
| h | human | 9606 |
| y | yeast | 4932 |
| d | Drosophila | 7227 |
| r | rat | 10116 |
| m | mouse | 10090 |

表1 首写字母与物种对应表

另外，统计了一下有多少摘要中我们完全没有找到物种名字出现，在总共644篇中，共有81篇。经过试验，无论加入还是不加入“hZIP2”这种蛋白名字中还有的物种，都是这81篇完全没有物种结果，所以他们都是属于在全文里才可能有物种信息的。

然后，我统计了一下各个物种在这个语料上出现的tf和df。整个语料上一共出现了74个物种，绝大多数都是只出现很少几次的，只有5个明显高于其他，如表2。可以看到，这5个物种就是表1中的那5个，这也验证了能用小写字母在蛋白质名字中表示的物种就是这5个。

|  |  |  |  |
| --- | --- | --- | --- |
| ID | 物种 | TF | DF |
| 9606 | human | 797 | 263 |
| 10090 | mouse | 773 | 246 |
| 4932 | yeast | 285 | 142 |
| 7227 | Drosophila | 176 | 104 |
| 10116 | rat | 58 | 37 |

表2 物种tf和df。这里的tf只是出现的次数，没有除以总词数。

对于字典中各个蛋白名字对应物种的数量做了统计，字典中共有358498个蛋白名字，其中277001个只有一种物种，81655个有多种物种，见表3。在只有一种物种的名字中，有276171个只对应一个ID，仅有830个有多个ID，说明主要的歧义确实是由物种引起的。

在字典全部名字中，共有84655个名字只有一个token，仅占所有名字的23.6%，而其中有33307个有多个物种，占全部物种歧义的40.7%，可见在只有一个token的名字中物种歧义占得比重比较大。

|  |  |
| --- | --- |
| 对应物种数 | 蛋白名字数 |
| 1 | 277001 |
| 2--10 | 73519 |
| 11--20 | 4198 |
| 21--30 | 1174 |
| 31--40 | 568 |
| 41—50 | 361 |
| >50 | 1835 |
| 全部 | 358498 |

表3 蛋白名字对应物种数统计