基于条件随机域的生物医学命名实体识别

李彦鹏，杨志豪，林鸿飞

（大连理工大学 计算机科学与工程系，大连 116024）

摘 要： 命名实体识别是生物医学文献文本挖掘重要的第一步。近年有很多人研究，然而效果并不理想。JNLPBA2004 测评中最好的系统只能达到 72.6%的 F-score。本文使用条件随机域(Conditional Random Fields，CRF)模型，采用 GENIA 语料进行训练，在 JNLPBA2004 测试集上得到了 71.9%的 F-score。本文讨论了不同规模训练语料，不同特征对 CRF 模型标注结果的影响。边界识别错误是识别中很严重的问题，本文针对左边界错误才采用了一种基于 CRF 的二次标注方法，使左边界错误率减少了 7.2%。

关键词： 命名实体识别；生物医学；文本挖掘；条件随机域

Recognizing Biomedical Named Entities using Conditional Random Fields

Li Yanpeng, Yang Zhihao, Lin Hongfei

(Department of Computer Science and Engineering, Dalian University of Technology , Dalian 116024)

Abstract: Named entity recognition is the first key stage of text mining in the exploding biomedical literature. While numerous approaches have been proposed, it remains a challenging task. The best performance in JNLPBA2004' task can only achieve an F-score of 72.6%. This paper presented a CRF (Conditional Random Fields) based system for identifying named entities using GENIA corpus as training set, and achieved an F-score of 71.9% in the JNLPBA2004'test set. This paper also discussed the effects of different size of training data and various features, and then proposed a two-stage tagging algorithm to resolve the left boundary errors. The results showed the left boundary error rate decreased by 7.2%.

Keywords: named entity recognition; biomedical; text mining; Conditional Random Fields

1 背景

随着生物医学技术的迅速发展，生物医学文献的数量也急剧增加。研究人员如何才能从海量的自然语言文本中获得所需信息呢？当今人们普遍采用文本挖掘(Text Mining)技术来解决这一问题。文本挖掘的第一步是命名实体识别(Named Entity Recognition，NER)。在生物医学领域 NER 工作比普通领域困难得多，JNLPBA2004 任务[1]的公开测评结果表明，在 GENIA[2]语料集上最好的系统也只能达到 72.6%的 F-score，离可以应用的水平还有很

基金资助：国家自然科学基金（60373095）

作者简介：李彦鹏(1981-)，男，辽宁，硕士，lyp_8218@163.com；杨志豪(1973-)，男，辽宁，讲师，博士，yangzh@dlut.edu.cn；林鸿飞(1962-)，男，辽宁，教授，hflin@dlut.edu.cn.

大的差距。

目前的生物医学命名实体识别的方法主要有基于字典和机器学习的方法。机器学习方法能够识别未登陆词，并且可以根据上下文环境对已经登陆词给出更准确的答案。因此越来越被人们所重视，大量的模型应用于该领域，。而其中最具优势的是既拥有马尔科夫链结构，又适合于处理复杂稀疏特征的条件随机域模型。从JNLPBA2004 测评的结果分析，系统[3]只使用了很少种类的特征，没有使用任何专业词典，F-score 就达到了 69.8%，而该实验使用的模型正是条件随机域。

2 条件随机域

条件随机域(Conditional Random Fields，CRF)[4]，是计算具有无向图 G 结构的随机变量集合 **S** 在给定随机变量集合 **O** 下的条件概率 $P(\mathbf{s}|\mathbf{o})$。

将 CRF 应用于命名实体识别中，则 **O** 表示一个句子的单词序列，**S** 表示相应的状态序列，标注的过程就是根据已知的单词序列推断出最有可能的状态序列，即 $P(\mathbf{s}|\mathbf{o})$的最大值。本实验使用了一阶线性 CRF。

$$P(\mathbf{s}\mid\mathbf{o})=\frac{1}{Z}\exp(\sum_{i}\sum_{k}\lambda_k f_k(s_{i-1},s_i,\mathbf{o},i)) \tag{1}$$

其中 $f_k(s_{i-1},s_i,\mathbf{o},i)$ 是二值特征函数，表明当前句子中第 i 个位置上是否具有第 k 个特征，并且取决于当前状态 s_i 和前一个状态 s_{i-1}。λ_k 是特征的权重，通过训练得到。

3 实验

3.1 特征选择

本实验借鉴了 JNLPBA2004 任务中各系统的部分特征，同时选取了一些新特征。共分为 9 类：

单词本身(F1)：将所有的单词都转化成小写字母。

构词特征(F2)：包括首字母大写，所有字母大写，是否包含横线，是否是数字等。

词缀特征(F3)：对每个单词都取了 3 个和 4 个字符的前缀，以及 3 个和 4 个字符的的后缀。

词形特征(F4)：将大写字母替换成A，小写字母替换成a，数字替换成0，特殊符号替换成x。

特征联合(F5)：将相邻位置的特征进行联合，得出新的特征，有助于识别长距离词。本实验选择窗口的大小为(-1,+1)。

词性标记特征(F6)和短语切分标记特征(F7)：本实验使用 GENIA Tagger 对训练语料和测试语料进行标注，得到相应的词性标记和短语切分标记作为特征。

关键词特征(F8)：实验中统计了训练集的命名实体中出现 20 次以上的 1-gram 和 2-gram 的关键词，将这些词是否出现作为特征。

边界词特征(F9)：从结果的统计中发现，相当多的错误都是发生在边界。因此，本实验统计了训练集中的边界词，取出现 5 次以上的作为特征。

3.2 结果分析

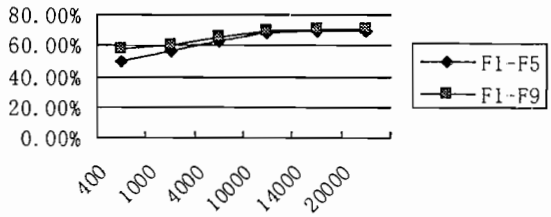

图 1 不同规模训练集的效果

Fig.1 Performance using different size of training data.

横轴表示训练集中的句子数，纵轴表示 F-score。曲线 F1-F5 表示采用特征 F1+F2+F3+F4+F5，F1-F9 表示采用全部特征。

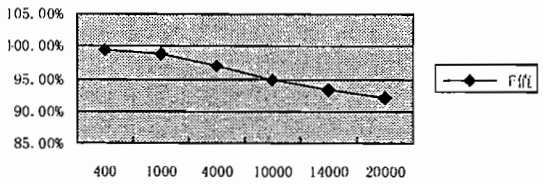

图 2 对训练集自身的标注结果

Fig.2 Performance on training data.

横轴表示训练集中的句子数，纵轴表示 F-score。

从图 1 可以看出随着训练语料的增加，后期的 F-score 趋于平缓，提高的幅度很小，一半的语料几乎没有被利用。可以预计，如果训练集再增加 2000 篇文章，还采用当前的方法，效果仍然不会有太大的改善。

另一个值得注意的现象是，随着训练语料的增多，对训练集本身的标注效果有明显的降低，见图 2。原因之一可能是由于有过多的特征是针对未登陆词的，从而影响了已经登陆词的识别；另一个原因则是语料本身的错误，尤其是标注不一致。有文献统计过生物医学文献人工标注的正确率在 87%-89%之间，但从本实验的结果推测，GENIA 中的标注正确率应高于 90%。此外，语料标注错误对机器学习方法的影响相对较小，统计的方法可以忽略极个别的错误；如果在训练集中的标注错误类型基本一致，只是通过学习错误的语料标注同样错误的数据，并不影响机器学习的效果。总之 70%左右的 F-score 不能仅仅用语料的错误来解释。但从曲线的趋势可以推测，仍使用当前的方法，无论使用多大的语料进行训练，F-score 都不会超过 92%。

3.3 边界判定问题

边界判定不准确是生物医学命名实体识别面临的最主要问题，经统计发现 38%的错误是发生在边界上。生物医学命名实体的边界判定是一个极其复杂的问题。本实验采用 CRF 进行二次标注，集中解决左边界问题。即固定了右边界，对左边届从新进行判断。首先把第一次标注的命名实体分为已确定和未确定的部分，左边界词以及它前面和后面一个词是未确定的，其它词都是已确定的。对已经确定的部分只赋予一个特征，就是第一次标注的结果，对未确定的部分赋予以下特征：

1. 词汇特征：1-gram，2-gram 项。
2. 关键词特征：与第一次标注相同。
3. 边界词特征：同上。
4. 词性标记，短语切分特征：同上。
5. 核心名词特征：假定右边界正确，把右边界词当作核心名词，如：purified human erythroid colony-forming

cells，cells 就作为核心名词。对于判断前面词的类别起着重要的作用，尤其当名字很长的时候。这个特征在第一次标注的时候由于没有判断出右边界而无法得到。

6．特征联合：将特征 1 的 1-gram 项和特征 5 联合。

然后将不同的结果利用 Google 进行裁决，仅仅利用简单的规则：如果较长的实体名长度不大于 3，且返回网页数超过 10 就算正确，否则选取较短的实体名；如果长度大于 3 且无不匹配的括号，则遵循第二次标注的结果，否则依照第一次结果。在 F-score 上得到了 1%的提高，左边界错误率减少了 7.2%（见表 4）。

表 3 二次标注的效果

Tab.2 Performance of the two-stage tagging.

实验编号	F-score（完全匹配）	F-score（右边界匹配）	F-score（左边界匹配）
一次标注	71.2%	78.6%	74.7%
二次分类	71.9%	78.6%	75.3%

4 总结

本文使用基于条件随机域的方法进行了生物医学命名实体识别的实验，讨论了训练语料规模和不同特征对标注结果的影响，然后使用二次标注的方法处理了边界判定的问题，取得了一定的效果。得出以下结论：对于通过一次机器学习很难处理的复杂的问题，往往可以利用第一次的结果选择新特征进行再学习，这样会逐步缩小范围，便于进一步处理。

参考文献：

[1] Kim JD, Tomoko O and Yoshimasa T, et al. Introduction to the Bio-Entity Recognition Task at JNLPBA. In the Proceedings of the International Workshop on Natural Language Processing in Biomedicine and its Applications (JNLPBA-04), 2004:70--75..

[2] Kim JD, Tomoko O, Yuka T, et al. GENIA corpus - a semantically annotated corpus for bio-textmining. Bioinformatics, 2003, 19(suppl. 1):i180-i182. Oxford University Press

[3] Settles B. Biomedical Named Entity Recognition Using Conditional Random Fields and Novel Feature Sets. In Proceedings of the Joint Workshop on Natural Language Processing in Biomedicine and its Applications (JNLPBA-2004), 2004:104-107.

[4] Lafferty J, McCallum A, Pereira F. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In Proceedings of the International Conference on Machine Learning, Williamstown, MA, USA, 2001.