关系抽取智能化语料标注系统 产品设计说明书 v0.1

1. 背景

自然语言处理的大部分任务是监督学习问题。序列标注问题如中文分词、命名实体识别，分类问题如关系识别、情感分析、意图分析等，均需要标注数据进行模型训练。深度学习大行其道的今天，基于深度学习的NLP模型更是数据饥渴。

最前沿的NLP技术往往首先针对英文语料。英文NLP的生态很好，针对不同有意思的问题都有不少大规模语料公开供大家研究，如斯坦福的 SQuAD 阅读理解语料。中文方面开源语料就少得多，各种英文NLP上的犀利模型和前沿技术都因为中文语料的匮乏很难迁移过来。另一方面，对于一些垂直领域，如医疗、金融、法律、公安等等，专有名词和特有需求甚多，很难将比较 general 的比如在 wikipedia dump 上面训练的模型直接拿过来用。

传统人工标注数据的过程往往是繁琐和低效率的。刚标了一个“联想”是公司名，又来一个“联想集团”，再标一次又来一个“联想集团有限公司”，如此的例子令标注过程含有大量的重复劳动。另一方面也没有一个易上手的标注UI，标注工作者往往需要直接按预先定好的格式直接在写字板之类的软件中修改原始数据，格式错误率也较高。

再者，典型的标注方法迫使项目进入不舒服的瀑布过程。项目只有在第一批注释完成之后才能开始，但注释团队在收到注解手册之后才能启动。为了制作注解手册，需要知道要尝试构建的功能需要哪些统计模型。机器学习是一项内在不确定的技术，但瀑布标注过程依赖于准确的前期规划，最终的结果是大量的浪费。

我们希望构建一个开源的中文文本标注工具，可以达到以下两个特点：

1. 标注过程背后含有智能算法，将人工重复劳动降到最低；
2. 标注界面显而易见地友好，让标注操作尽可能简便和符合直觉。

> 参考资料：
> https://github.com/crownpku/Chinese-Annotator
> https://prodi.gy/docs/

2. 技术方案

在标记难以自动获取的情况下，通常需要由领域专家进行人工标记。让专家对大量实例进行手动标记的过程是极为不经济的，我们期望在标记尽可能少实例的同时达到较高的预测准确率。从直觉上来看，随机地选择实例给专家标记并非最佳策略，我们希望学习算法可以主动地提出一些标注请求，将一些经过筛选的数据提交给专家进行标注。

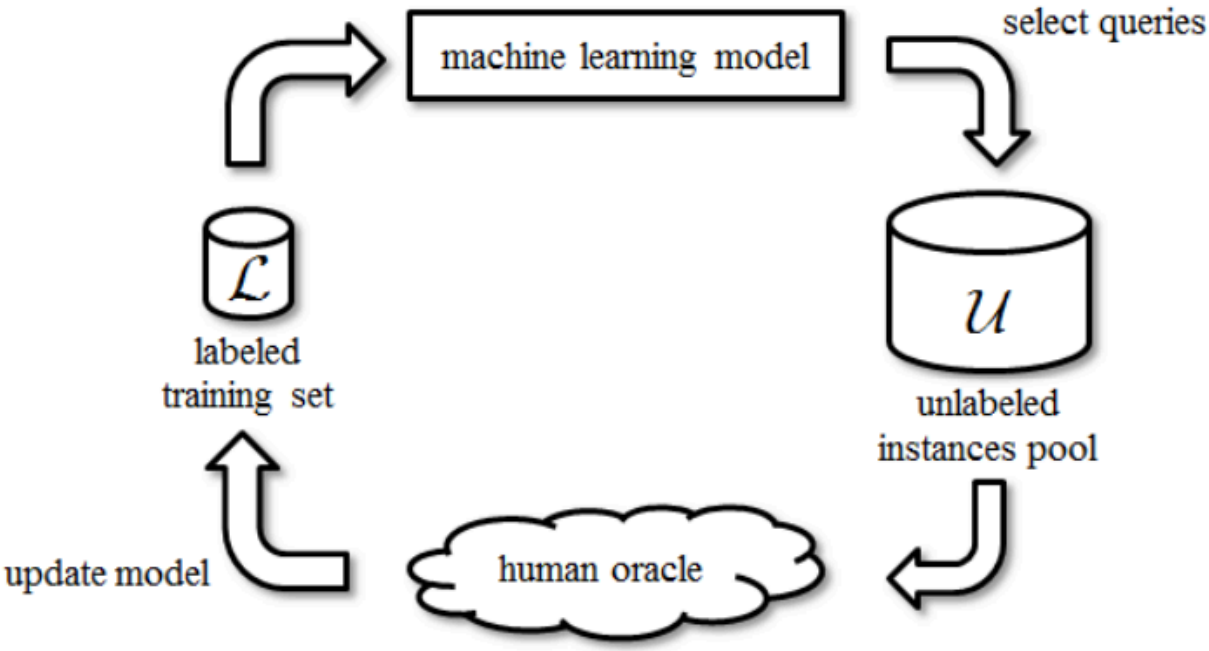

Fig.1 Active learning process

图 1 主动学习过程

这就是主动学习算法，它是一种统计学习方法，其思想来源于统计学中的最佳实验设计。

假设在统计学习算法的运行过程中，已标记实例集为 L，未标记实例集为 U，则主动学习的任务就是设计对实例“价值”的评估函数 f，其能够在未标记的实例中选择一个最佳的实例 u* = argmax_u∈U f(u,L,U)，对其标记后能够最大程度地提高模型的预测准确率。通常，主动学习算法采用启发式的贪心策略，即每次从未标记实例中选择某一方面属性最大(或最小)的实例进行标记。

下面是我们要考虑的选择策略：

1. 随机采样 RND
   从未标记的实例集 U 中随机选择一个 x 进行标记

2. 不确定性采样 Uncertainty Sampling (US)
   这是一个支持向量机算法：总是选取距离分平面最近的实例作为不确定性最大的实例进行标记。有研究表明，它不仅是一种典型的 uncertainty sampling 算法，还是一种最小化解释空间的启发式算法。其原理简单，时间复杂度低，准确率高。

> 参考资料：
> http://lamda.nju.edu.cn/huangsj/dm11/files/jiangyy.pdf 主动学习在未标记数据挖掘中的应用 蒋炎岩 南京大学 计算机科学与技术系, 南京 210046
> http://www.10tiao.com/html/600/201610/2652550384/1.html 主动学习 Active Learning