微博客蕴含交通事件信息信息抽取的自动化标注方法

最先阅读的是仇培元，张恒才等的《微博客蕴含交通事件信息信息抽取的自动化标注方法》一文，通过微博客中获取的丰富的交通事件信息，为现有的交通信息收集手段提供补充。

在前人基础之上主要解决问题： 以往缺少对地理实体关系的判断，对涉及多个地理实体及关系表达的地理空间要素抽取效果不佳的问题

解决方案： 将地理实体关系识别引入事件抽取的过程中来。 利用CRF模型实现交通事件角色标注，在利用SVM实现角色关系与要素关系的标注，来完成交通事件信息空间要素识别

交通事件主要由 1空间要素2时间要素3主题要素 组成

A、交通事件角色标注：：：

角色标注问题 可以视作为 序列标注问题，解决序列标注问题常用最大熵，HMM，CRF等。CRF没有严格的独立性假设，可以利用上下文信息，计算结果为全局最优，克服序列标注偏置的问题，在序列标注的问题上效果最优

对于机器学习模型来讲，特征选择十分重要---》

特征选择：（特征为词和词性对） 自身特征（当前第i词及词性），相邻词特征（第i-2，i-1，i+1,i+2词及词性），组合特征（第i-2和i-1词组合及词性组合，第i-1和i词组合及词性组合，第i和i+1词组合及词性组合，第i+1和i+2词组合及词性组合，第i-2和i-1和i词组合及词性组合，第i-1和i和i+1词组合及词性组合，第i和i+1和i+2词组合及词性组合）

B、角色语义关系标注：：：

解决问题：网络文本中蕴含多个交通事件的时候，其信息数量事先无法预知， 在事件角色标注的基础之上，使用机器学习算法标注角色之间的语义关系，将具有语义联系的角色实例构成实例关系网，从而完成隶属于不同交通事件的事件角色

为实现空间要素提取，首先选取出空间要素的角色实例 ，两两组合成候选二元角色对，并过滤掉类型组合无意义的角色对

为判断角色对之间的二元关系是否成立，及“是”与“否”的二元分类问题 训练数据有限，特征选择相对丰富，SVM模型的优势在于采用结构风险最小化原则和函数思想，在小样本、非线性及高维模式识别中效果较好

特征选择： 词及词性组成的对---自身特征，相邻词特征，特殊特征（i词和j词的位置关系，等）

经角色关系标注之后得到具有语义联系的角色实例对集合， ------>构建角色实例网 ； 选择具有起点作用的角色 开始，从实例网中划分出子网，每个子网实例构成某一事件的空间要素

C、然后获取和空间要素相关联的主题要素：：：

步骤 -筛选出与主题要素描述相关的角色实例，通过交通事件词典cast到标准的主题要素集合 再使用SVM将具有语义关联的空间要素和主题要素的组合

特征选择：空间要素角色特征、主题要素角色特征以及一些特殊特征

D、事件时间要素抽取：：：

正则表达式匹配，时间的格式比较稳定

=====================================================================

综述：中分分词十年回顾

精读了《中分分词十年回顾》（这十年间主要进展了基于统计的机器学习方法，未引入神经网络）

主要讲述了97年到07年十年间的中分分词技术的变更和进步，最开始“分词规范+词表”的组合被认为能够界定句子中的词语，评测中掺杂了人的主观判断，缺少对什么是词的可计算定义，03年SIG汉举办的中分分词评测大赛Bakeoff ，引入了严格的质量控制，就是特定的分词语料库， 分词规范和词表在制作高质量的分词语料库中发挥着作用。

还有重要的一点就是中分分词所依据的只是文本的表层信息。中间掺杂着 理解与分词 谁先谁后的思考，最终通过SIG汉的中分分词大赛的不断的进行，最终被验证以分词为先可以在分词系统上具有更好的效果。

一直以来，未登录词和分词歧义一直是影响中文分词的最重要的两个因素，经过验证，在语料库足够大的情况下，未登录词对分词精度造成的影响比失落歧义造成的影响至少大10倍左右，因此能大幅度提升对未登录词的分词精度的分词方法也能够提升分词系统的总体性能。

基于字标注 的分词方法实际是构词方法，即把分词看作是字在字符串中的标注问题，因此它能够平衡的看待词表词和未登录词的识别问题。随着基于字标注的分词类型被验证拥有更好的效果，其中基于字标注的常用标注方法有最大熵，HMM，CRF模型。 其中，分词在中文信息处理中的初级地位，因此可供选择的特征非常少，一般为字本身 以及 词义转移概率。CRF模型更只需要设置字特征，词位学习中确定字特征的主要参数是上下文窗口的宽度。其中条件随机场训练分词模型时，要统筹选择词位标注集和特征模板集，通常这样能获取更好的效果

=====================================================================

基于Web新闻的时间抽取与时空分析系统

《基于Web新闻的时间抽取与时空分析系统》 主要讲解了使用网络上的开源工具对整个项目的流程进行介绍，其任务具有特殊性，即一条文本只具有一个特定事件，属于被预处理特别好的文本，具体的工具包给读者提供借鉴意义。

=====================================================================

阅读了博客上一个 ID名为 我爱自然语言处理，大家成为52nlp的大神的帖子，关于一些常用语料库，多为SIG汉的比赛使用的语料包。阅读了他的关于中分分词的5、6篇博文，很是经典。然后下面是我参考的博主的文章，并进行实际操作的两页博客，使用张乐博士的最大熵工具包，以及一个开源的CRF++的切词工具包在ubantu系统上的linux环境下进行实际操作

http://www.52nlp.cn/%E4%B8%AD%E6%96%87%E5%88%86%E8%AF%8D%E5%85%A5%E9%97%A8%E4%B9%8B%E5%AD%97%E6%A0%87%E6%B3%A8%E6%B3%954

http://www.52nlp.cn/%E4%B8%AD%E6%96%87%E5%88%86%E8%AF%8D%E5%85%A5%E9%97%A8%E4%B9%8B%E5%AD%97%E6%A0%87%E6%B3%A8%E6%B3%953

======================================================================================================

======================================================================================================

最近在阅读其他老师推荐的论文，正在阅读《神经网络事件抽取技术综述》，截止当前，我对神经网络下的中分分词还未理解，最近的工作是多阅读这方面的论文及综述。

------------------------------

Event Detection and Domain Adaptationwith Convolutional Neural Networks

论文提出之前的各种方法的不足之处：

1、特征的选择任然是一个手动的过程，需要语言的直觉以及各个应用域的专家的指导

2、使用了NLP工具的系统，会导致这些工具在工作工程中的错误传播

本文提出了一种新的域自适应的CNN事件检测框架来减轻上面提出的两种不足，

事件检测视为多分类问题--》

检测句子中的每个token是否为触发词，当前token和它的上下文组成候选触发词或者例子

将句子进行修剪或者填充的预处理为固定长度的句子，方便喂给CNNs

句子中的每个token都被处理为 词向量矩阵中的词嵌入表示+位置嵌入+实体类型嵌入

网络结构： 卷积层+最大池化层+softmax

将a域数据作为资源域来训练模型，用b域作为目标域来测试CNNs在域适应方面的能力

--------------------------------------------------------