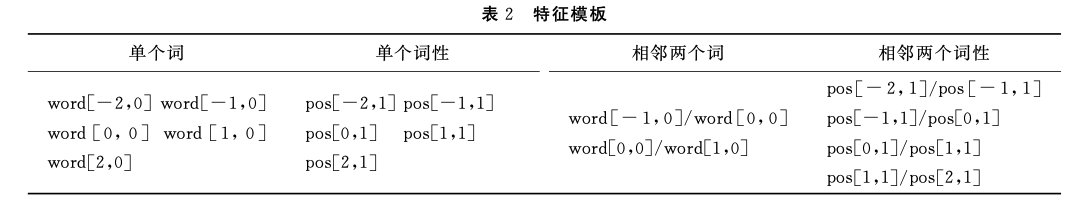
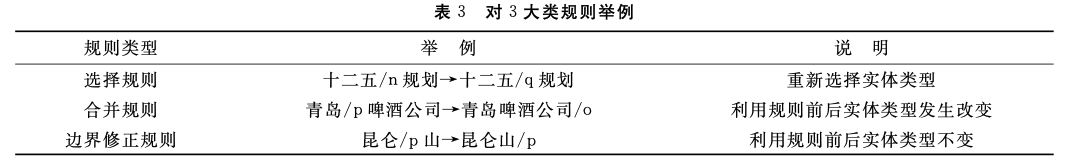
**论文阅读笔记一**

基于CRF的互联网文本命名实体识别研究

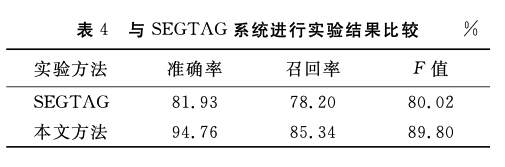
1. 主要贡献
2. 引入中文实体匹配规则对CRF模型的识别结构进行补充修正。
3. 探究对比了规范化和未规范化的文本的实体识别效果。
4. 方法
5. 首先用CRF模型进行实体识别。
6. 设置CRF模型参数，包括L2正则化算法a、拟合训练数据程度值c、特征出现次数f、多线程p。
7. 特征模板的设置，主要考虑词和词性。



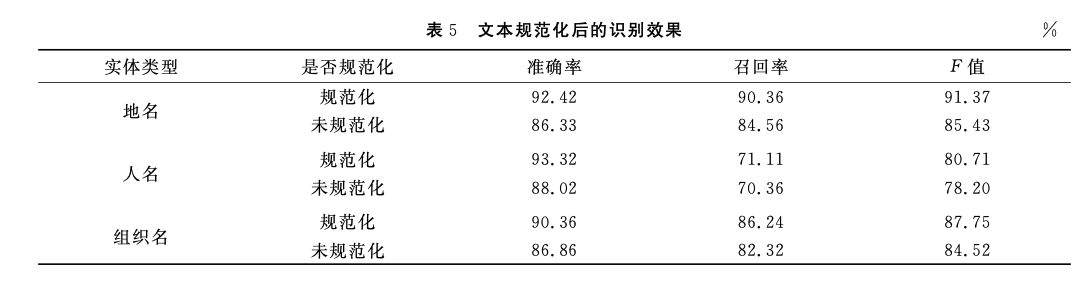
1. 语料集处理。对人名、地名和机构名进行整体标记。
2. 然后结合实体匹配规则对CRF的结果进行修正。



1. 实验结果分析
2. 实验语料：哈工大智能技术与自然语言处理实验室提供的2013年的报纸新闻和网络新闻，总计160万个词。未排除发布时间导致的实体重复率过高故本实验选取8000篇新闻作为语料。
3. 评价标准：P、R、F
4. 与SEGTAG系统的对比

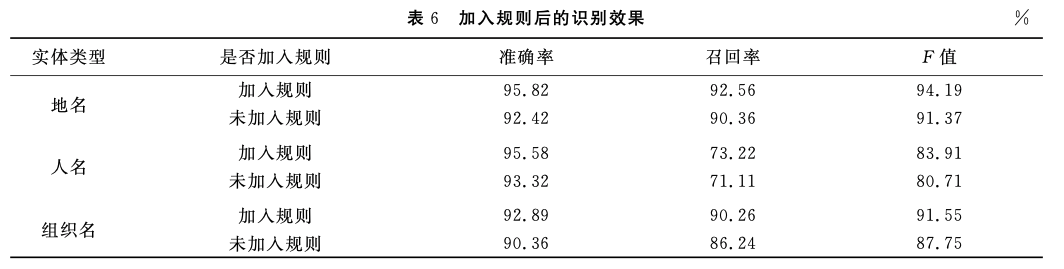


1. 文本规范化对实体识别的影响



实验表明，文本规范化对命名实体识别很重要。人名、地名和组织名均由名词或形容词组成，去除其他词性的词之后，实体识别的效果有很大的提高。

1. 加入规则后的识别效果



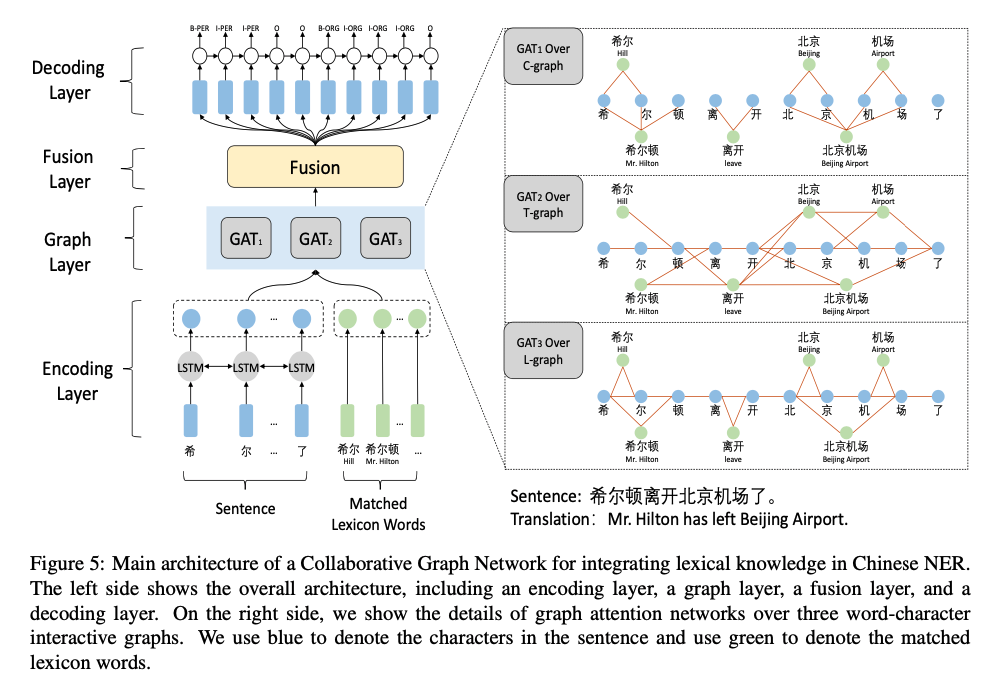
规则的引入对第一次识别后得到的结果进行了修正，获取更加准确的结果。实践中可以结合文本的特点引入上述规则（选择、合并和边界修正）

**论文阅读笔记二**

Leverage Lexical Knowledge for Chinese Named Entity Recognition via Collaborative Graph Network

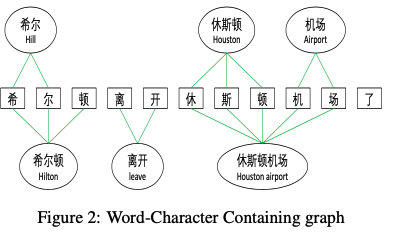
1. 主要贡献
2. 作者提出了一个协作图网络，该网络高效直接地整合词汇信息进行中文实体识别。
3. 该方法提出了三个单词特征交互图用于解决单词匹配问题和最近上下文词汇选择问题。
4. 方法
5. 模型结构

模型包括编码层、图形层、融合层和解码层。

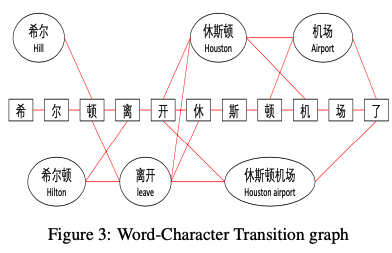


1. 图形层介绍

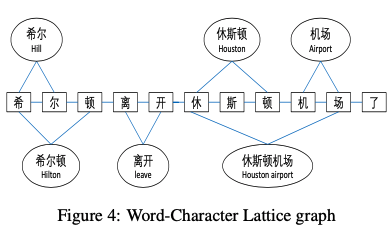
图形层包括三个部分，第一个是单词-字符包含图C-graph，该层为嵌入自匹配词汇设计的，模拟字符和模型匹配出的单词的联系；如果词汇i包含字符j，则C图对应的邻接矩阵的(i,j)项值为1。



第二个是转移图单词-字符转移图T-graph，该图建立了字符和最接近上下文匹配词之间的直接连接，有助于直接集成最近的上下文单词；如果一个词i或字符m与最接近的前一个或后一个子序列匹配，则T-graph的邻接矩阵 (i,j)或(m,j)项值为1。此外，为了获取词汇之间的上下文关系，如果词汇i是另一个词汇k的前后上下文，我们将置为1。



第三个是单词-字符格子图L-graph，通过多次跳跃隐式获取自匹配词汇词和上下文词汇的部分信息；如果字符m是字符j的最接近的前一个或后一个字符，则L-graph对应的邻接矩阵的(m,j)值为1，如果字符j与单词i的第一个字符或结尾字符匹配，则将置为1。



1. 协作图网络模型

模型包括编码层、图形层、融合层和解码层。

1. 编码层：捕获句子的上下文信息并表示词汇的语义信息。

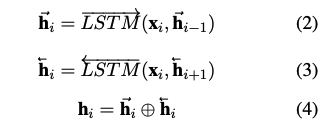
输入句子被表示为s = {c1，c2，...，cn}，其中ci表示第i个字符；

匹配得到的词汇为l = {l1，l2，...，lm}，从预训练字符向量表中查找字符的嵌入，得到每个字符的向量表示：



其中是字符词嵌入查找表。

模型用BiLSTM捕获上下文信息：



从预训练单词词嵌入表中查找单词的词向量：



其中是单词词嵌入查找表。

将上下文信息和单词词嵌入进行拼接，得到每个节点的向量表示：



1. 图形层：基于GAT模型，对C-graph、T-graph和L-graph进行建模。

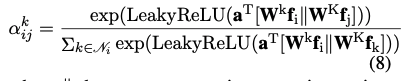
单词-字符交互图上的图注意力网络：

在一个M层的图注意力网络GAT上，第j层的输入是一组节点特征以及邻接矩阵A，，其中，N表示节点数，F表示第j层要素的维数。第j层的输出是一组新的节点特征：



具有K个独立注意力头的GAT运算可表示为：



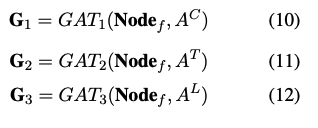


其中||表示拼接运算，σ是一个非线性激活函数，是图中节点i的邻居节点，α是注意力层的系数，W和a是单层前馈神经网络。

最后输出层的特征为：



独立构建三个不同的单词-字符交互图，这三个图贡献相同的顶点集，所有节点的输入都是，输出节点特征如下：



仅保留这些矩阵的前n列，丢弃最后的m列，仅使用字符来表示解码标签：



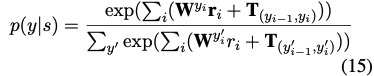
1. 融合层：融合图形层获取的词汇知识。

融合层的输入是上下文表示H和前一层的图形层输出Qi：

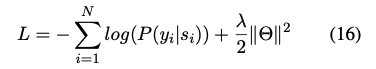


其中，参数W是待训练的参数，本层输出R是新的句子表示形式，融合了上下文信息以及自匹配词汇词和他的上下文词汇词的词汇知识。

1. 解码层：使用标准的CRF模型对标签进行解码。



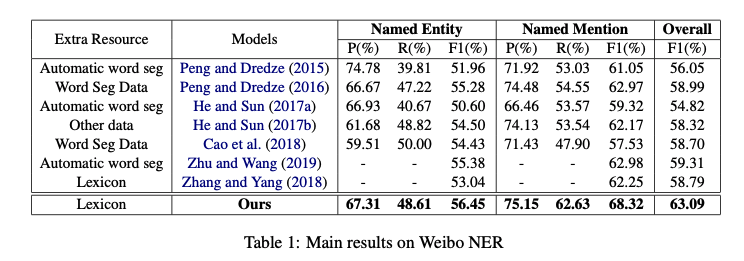
损失函数：

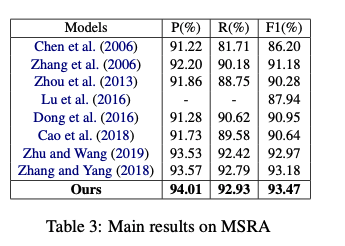
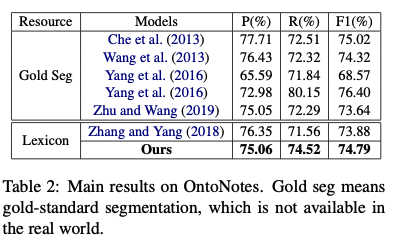


λ是L2正则化的参数，θ是待训练的参数集。

1. 实验结果和分析

数据集：社交领域weiboNER、新闻领域OntoNotes和MSRA





在OntoNotes数据集上，作者提出的方法不依赖于数据集的分割情况。

在MSRA数据集上，作者提出的方法也优于之前的模型。

在模型性能方面，作者提出的模型可以通过批量训练进行计算，同样条件下模型比LatticeLSTM快6-15倍。