首届AIIA杯人工智能巡回赛
国家电网站决赛

电力领域专业词汇挖掘

答辩人：刘辉

目录》

CONTENTS

01

PART

首届AIIA杯人工智能巡回赛国家电网站决赛

研究背景与意义

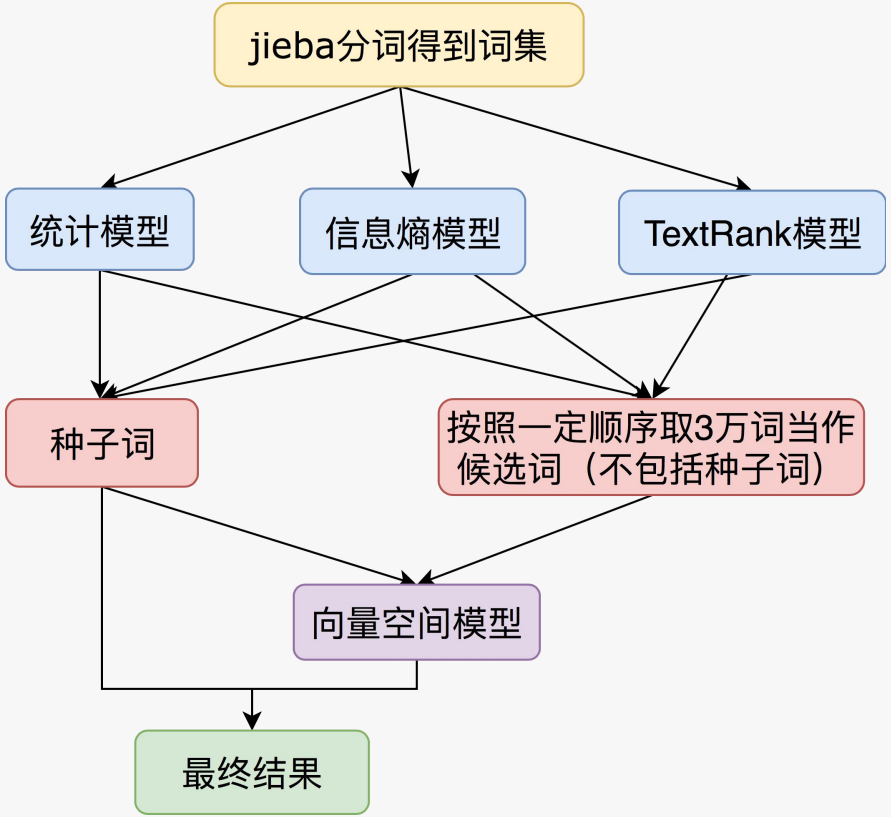
- 目前电力行业还没有建立较完善的电力主题词典，使用电力主题词典支持相关文本语义理解的需求不断增加。
- 电力行业已积累了大量的文本数据，包括电力科技论文、项目报告、电力规程、电力操作手册等。人工筛选电力专业领域词汇效率低下，而利用自然语言处理技术可以快速、高效地发现电力专业领域词汇，构建电力主题词典。

02
PART

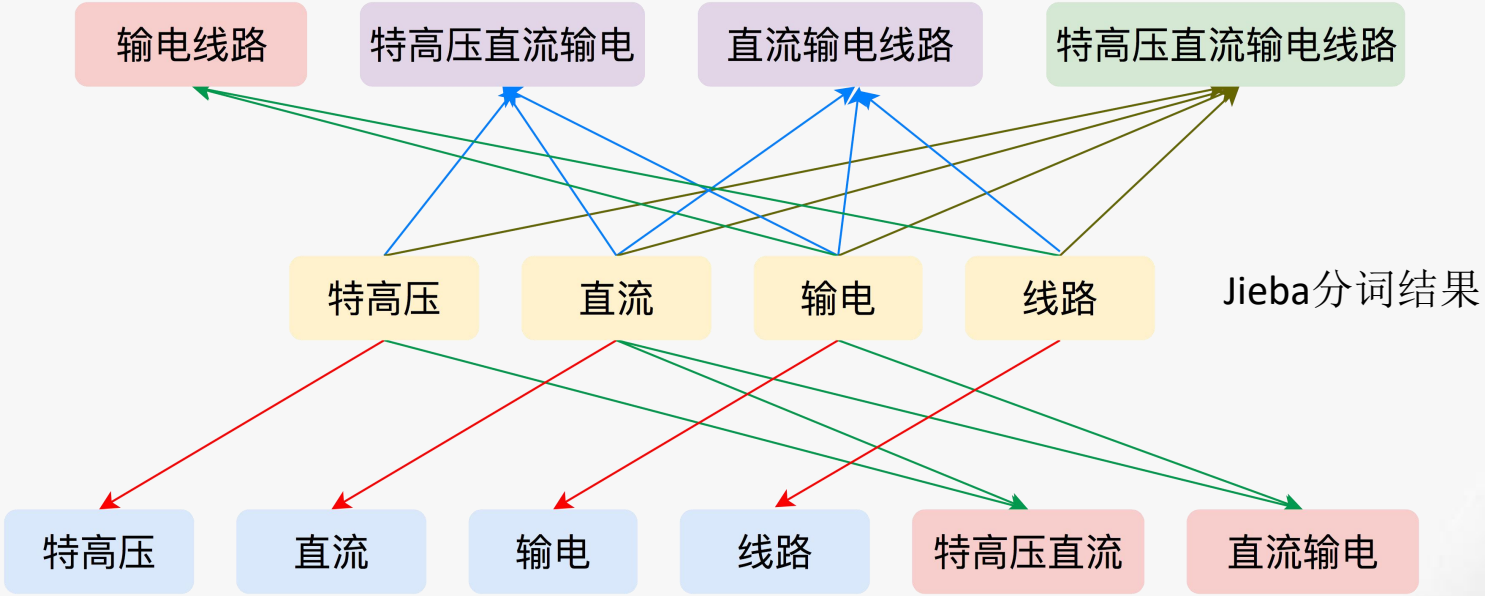
首届AIIA杯人工智能巡回赛国家电网站决赛

研究方法与思路

模型框架


jieba分词得到词集
统计模型
信息熵模型
TextRank模型
种子词
按照一定顺序取3万词当作候选词（不包括种子词）
向量空间模型
最终结果

- 直接使用Jieba进行分词的粒度太小，难以用来寻找专业词
- 将Jieba分词结果相邻的1个、2个、3个、4个词进行组合生成新词



分词结果过滤

- 停用词表过滤：如果分词结果中出现了停用词，如“的”，“和”等，直接进行过滤
- 改进的互信息过滤：$P(*s)=\frac{tf_{*s}}{tf_s}, P(s*)=\frac{tf_{s*}}{tf_s}$

  其中tf_{*s}表示*s去重后的频度，tf_{s*}为s*去重后的频度

提取局部放 [-1, '来'] ['电', '电'] → P(*s)=1; P(s*)=0.5

局部放电脉冲 ['取', '取', '于', '段']['特', '的', '在', '波'] → P(*s)=0.75; P(s*)=1

放电脉冲特征 ['部'] ['的'] → P(*s)=1; P(s*)=1

统计模型

- 统计模型使用改进的TF-IDF模型作为评价标准
- 改进后的模型加大了DF的惩罚，实验证明效果很好

原始TF-IDF：$w_i = tf_i \times \log\left(\frac{N}{df_i}\right)$

改进TF-IDF：$w_i = tf_i \times \log\left(\frac{N}{df_i+20}\right)$

信息熵模型

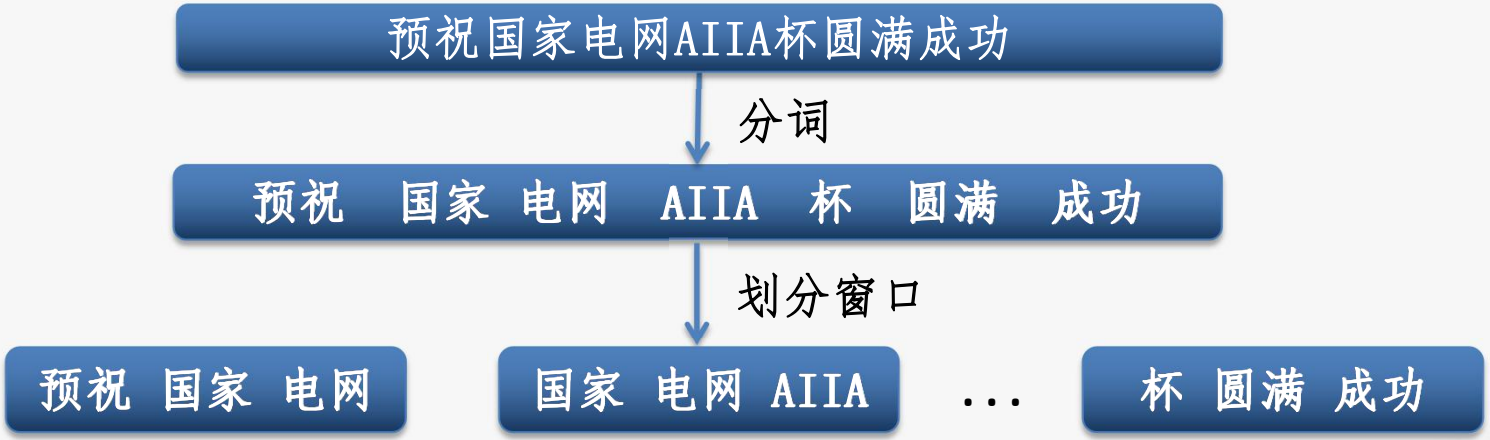
- 信息熵模型使用左右信息熵的乘积作为评价标准
- 关键词的左右能搭配的词丰富，左右信息熵的乘积大
- 信息熵$H(U) = -\sum_i p_i \log p_i$

提取局部放 [-1, '来'] ['电', '电'] → $H(w)=H(w_{left})*H(w_{right})=1*0=0$

局部放电脉冲 ['取', '取', '于', '段']['特', '的', '在', '波'] → $H(w)=H(w_{left})*H(w_{right})=3/2*2=3$

放电脉冲特征 ['部'] ['的'] → $H(w)=H(w_{left})*H(w_{right})=0*0=0$

- TextRank模型将词视为“节点”，构建出词关系图，根据词之间的共现关系计算每个词的重要性
- 当两个词出现在同一窗口中时，词关系图中就会有对应的连接边



TextRank模型

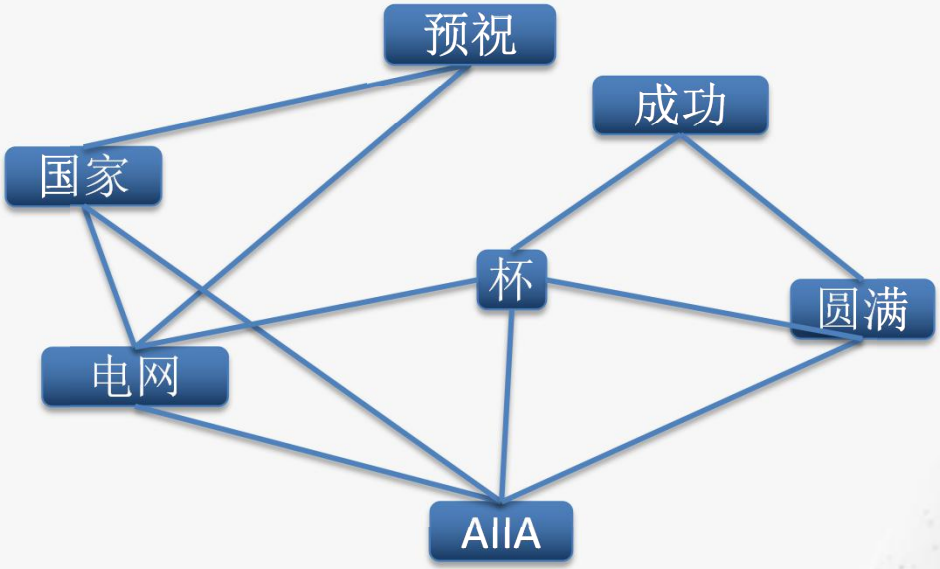
预祝 国家 电网

国家 电网 AIIA

...

杯 圆满 成功

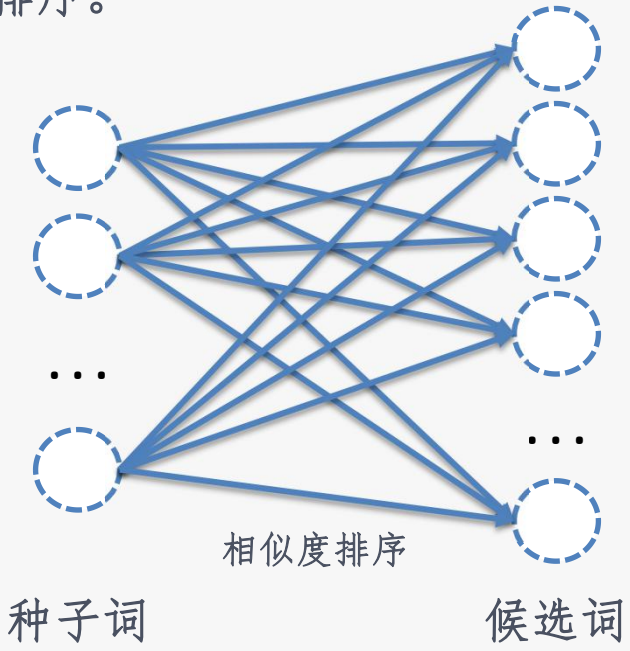
构成网络



$$\text{TextRank: } WS(V_i) = (1-d) + d * \sum_{j \in In(V_i)} \frac{w_{ji}}{\sum_{V_k \in Out(V_j)} w_{jk}} WS(V_i)$$

- 使用上述三个模型的结果（3w专业词）作为Jieba词典，重新对训练集数据进行分词，使用word2vec模型计算得到各个词的空间向量
- 使用上述三个模型结果的交集作为种子词集，剩余其他的词作为候选词集

- 对于每个种子词，在候选词集中寻找相似度最高的k个词，然后投票并按照投票结果进行排序。



03

PART

首届AIIA杯人工智能巡回赛国家电网站决赛

总结与展望

- 提出了一种基于种子词与候选词相似度投票的专业词筛选方法，效果表现良好。
- 模型框架简单，容易实现，可以作为其他相关研究的基础。
- 模型主要时间花费在分词处理和word2vec中，总体效率高。
- 模型没有使用电力领域知识，理论上具有良好的泛化能力，除了电力领域，还可以用于其他领域词汇挖掘。
- 模型未使用任何标注数据，是一种纯无监督模型。

- 当训练集增加或者有更多更准确的种子词集时，理论上效果会更好。
- 向量空间模型中如果加入更多特征，理论上效果可能更好。
- 模型可以用于专业领域词汇挖掘、新词发现、关键词挖掘、相似主题挖掘等领域。

感 谢 聆 听!

刘辉
liuhui@iie.ac.cn