**电力词典新词识别引擎的设计与实现**

**背景：**

源于《“AIIA”杯-国家电网-电力专业领域词汇挖掘》竞赛项目。以电力专业领域知识为例：词语是开展文本处理的基础，目前电力行业还没有建立较全的电力主题词典。随着对电力文本语义理解需求的增多，构建电力行业主题词典需求越来越迫切。电力行业积累了大量的文本数据，这些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。基于这些数据，利用自然语言处理技术开展电力专业领域词汇发现研究，进而构建电力主题词典，对于后续开展电力行业文本理解和挖掘具有重要的意义。

原型系统的任务：对给定的电力文本数据，利用专业领域词发现算法来挖掘电力专业词汇。参赛者需要根据提供的电力语料，对比通用语料（建议采用维基百科），开发专业词发现程序，寻找出专属于电力领域的词汇。电力语料格式为一个Txt文本集，该文本包含了10000篇电力科技论文（已打乱顺序）。

**关键词：电力词典 新词识别 模型 引擎**

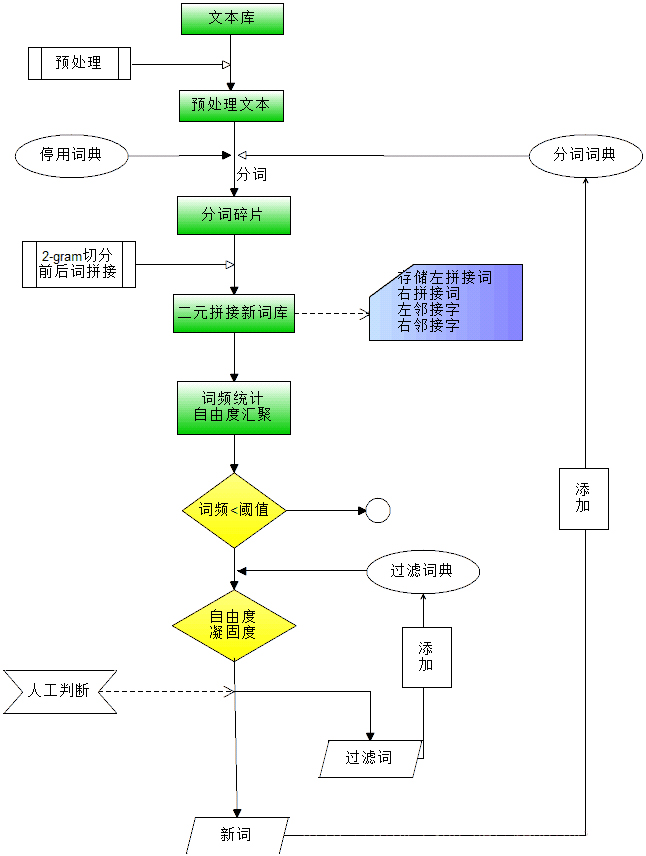
**基本架构：**

整体架构由若干流程构成，每一个流程对应一个模块，每一个模块对应一个或几个算法。  
设计和理顺架构图的目的是：要想提高最终分析结果，就可能要对模块进行优化或调整，流程的层进关系必须清晰，最终抽象为一个架构图。



三大模块：

（一）语料预处理（清理）+ 工具准备（专业词典，过滤字典（生活常见词，停用词，散碎符号表））  
（二）待选词表生成（新词发现）



新词识别的一般型处理流程示意图

（三）挖掘领域词(相似度模型)

**数据预处理：**

**主要处理流程及数据挖掘模型算法说明：**

主要分三大类的方案：

4-1 统计发现法：主要用于模块二，即生成候选词集，广义上说字典匹配法也属于统计发现法，因为字典的建立也基于统计的方法，一个人认可为词，多人认可为词，即可认为是字典中的词。

4-2 构词法：主要用于模块三，即挖掘领域词。

4-3 综合（融合）法：对于两个处理模块发现词与筛选词都适用。

4-1-1 对词表的词元素先进行TF\_IDF排序，然后用传统语言模型进行多元词元素组合，发现新词，选入新词队列。

4-2-1 建立word2vec模型，然后按照标准字典，进行近义词处理，选排名前3或5的词，进入候选新词队列。

4-2-2 半监督深度学习模型，将建立的词向量队列同现有专用字典匹配，用全部匹配词和2-3倍（此倍数可以由随机抽取语料统计得出）的非匹配词构建训练集与测试集。单字符为特征，训练文本分类深度学习模型：CNN+bilstm二元分类器（匹配专用字典为1 不匹配 为0），剩余没有选入训练集与测试集的词向量队列，用训练好的分类器进行分类，将识别为新词的队列与已经命中于现有字典的队列，构建最终的结果识别队列。

4-2-3利用传统的命名实体识别方法：例如规则（决策树法）：含“电”，“器”，“机”，“组合”，“阻尼”等字在行区分的句向量，进行最后校验或作为基线。

4-3-1 针对4-1-x和 4-2-x处理子模块，建立传统融合分类模型（如xgboost） 。子模块存在 得1分，不存在得0分，抽取一部分做训练数据，然后二分类形成最后的结果识别队列。

**数据分析结果测试:**

**软件工程测试：**

**改进要点：**

**总结：**