原理是这样 从字典 和 第一模型融合 找交集的 若干高频词 这高频词 几乎在测试榜单 是100%命中 然后 再相似度 用的是word2vec 然后成绩就上去了我觉得 利用了 语言模型的预测本质：也是一种对共现统计特征的学习。相似模型里 w2v比较成熟 所以 只要构词法种子词 选得好 语料专业性强（领域集中） 那一定能构词法种子词 来源于 前面基于最大熵原理和teaxtrank的模型融合 本身具有 跟语料强专业语义耦合性 和 历史专业词典 做交集 历史专业字典 本身固有的词 都是凝结了专业智慧的人类智能 都是经过专业领域经验构词法处理过的挖掘出很多该领域的新词 就是说 字典的词 都是 构词法构造出来 共现统计的因素很少 而最大熵和textbank 的融合构造的候选词集 明显是 共现统计规律为主导 二者取交集 那么 就是 即满足构词法 又满足 统计共现

同理 满足专业构词法的词 一定是专业词 这个是公理

那么 我用深度学习得来跟 构词法统计词 近似的候选词集里的 结果词集 满足构词法成词的概率很大

正向推理： 满足于构词法得词是专业词 》》》 大概率满足构词法的候选 是专业词

再解释下 我简单粗暴的解释 我前面得到候选词 是满足共现统计的 如果挑出来 候选集 里 跟字典有交集的词 做一个语料的w2v近似词模型，表达下隐含的语义相关 再返回找 候选集里没在字典出现 但是和种子词语义相似的词 那这些词 即满足 统计共线规律 又 大概率满足专业构词法 那一定是大概率专业词 答案得出

专业构词法： 只有专业人士通过长期的领域知识学习和经验总结，对词根进行共现组合表达出专业领域知识点的有别于其它领域和日常词的词根组合。

我这么理解 希望答辩的时候 能讲清楚 小光说的算法 是满足 共现统计规律 基本原理 最大熵的 算法 目的是 找出 大概率的候选词集 其实不让用字典直接匹配 大家弄来弄去 搞的都是这个东西

后期排行 比的是 怎么“提纯”是吧？提纯涉及两个方面 一个是查全 候选队列里有多少个 专业词 一个是查准 怎么用内隐的专业构词法规则 挑选专业词

我这个又一个失误 就是 如果语料 基本全是专业文章 那么 纯粹的统计规律 绕过构词法的效率到底有多强 抛开过滤方法