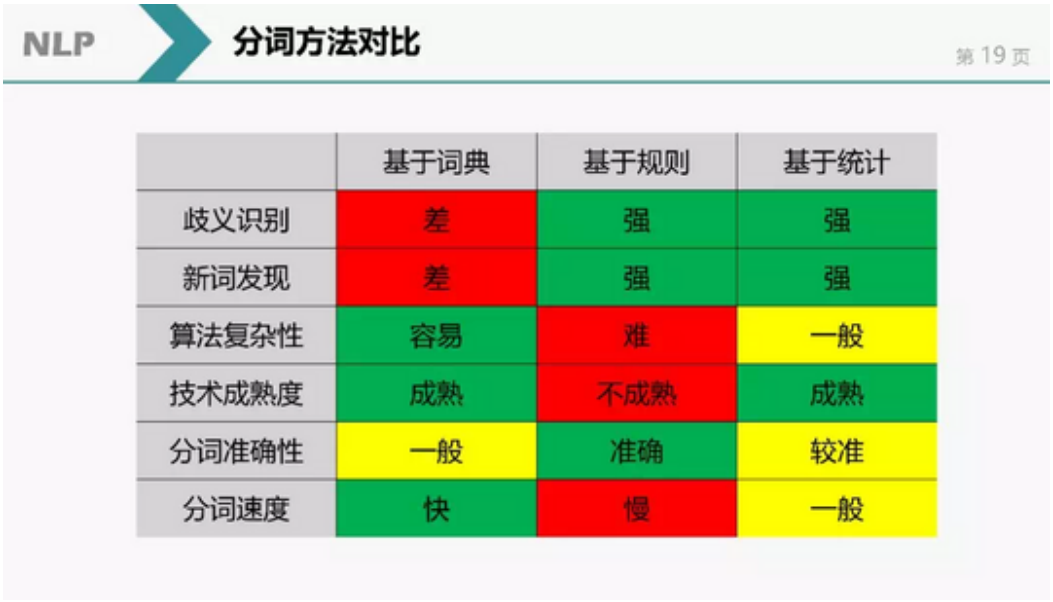
NLP有意思的问题汇总、典型的NLP例子

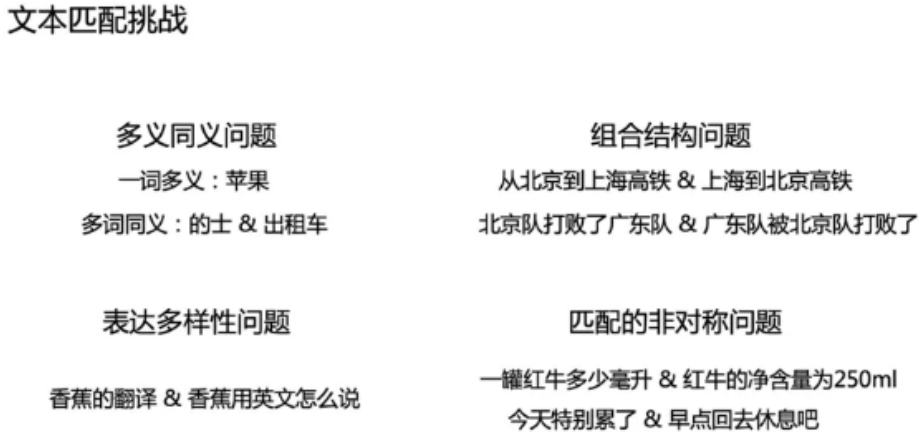
笔记本: 自然语言处理
创建时间: 2019/9/4 15:02 **更新时间:** 2019/9/9 9:10
作者: beyourselfwb@163.com

来源： https://www.codercto.com/a/66150.html



	基于词典	基于规则	基于统计
歧义识别	差	强	强
新词发现	差	强	强
算法复杂性	容易	难	一般
技术成熟度	成熟	不成熟	成熟
分词准确性	一般	准确	较准
分词速度	快	慢	一般

来源：
https://mp.weixin.qq.com/s/BdvBV542AZnMw7hnjDSwWw



参考：https://cn.100offer.com/blog/posts/296

NLP 算法工程师的学习、成长和实战经验

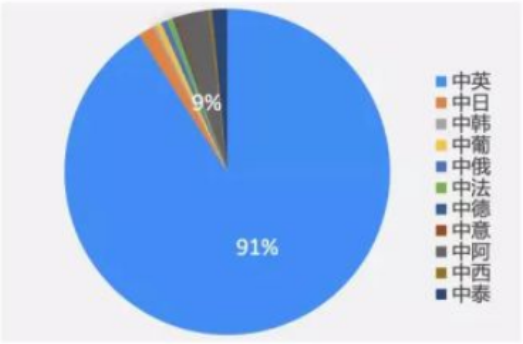
NLP解决的**五个基本问题**，这五个是李航老师在北大的AI公开课上提出来的，分别为：

1.分类问题：分类问题大家平时表容易见到，比如文本分类，情感分析目的是把一段文本打上一个或多个标签。
2.匹配问题：比较常见的是检索，检索与某句话类似的话或者是与它相关的回答，这个就是匹配。
3.翻译问题：类似于两种语言之间的翻译，把一种语言翻译成跟它语义相似的另外一种。
4.结构化预测：把一段文本结构化，所谓的结构化类比于一段文本中的词对应的是动词还是名词，语法角色是主语还是谓语，将其转化为结构化的输出序列。
5.马氏（马尔可夫）决策过程：当前要采取的动作和上一个状态和动作相关，这是一个典型的马氏链的过程。代表性的系统是对话，如何回复当前用户的话，是和最近的上下文相关的，这就是马氏决策过程。

机器翻译存在的问题和挑战
来源： https://www.zhihu.com/question/24588198

机器翻译挑战 – 数据稀疏

人类语言超过5000种



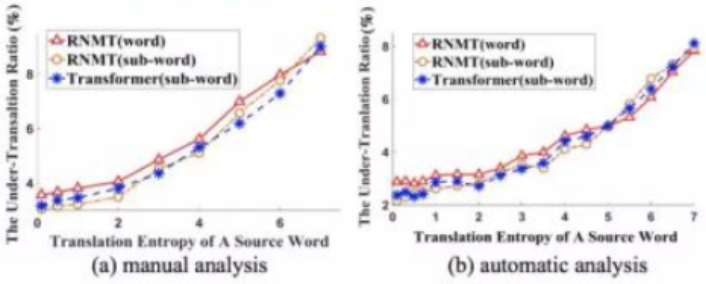
中文相关主要语种双语资源分布

挑战一：漏译

Source: 最终 真 善 美 彻底 打败 了 假 恶 丑

漏译与词语的熵成正相关

数据方面

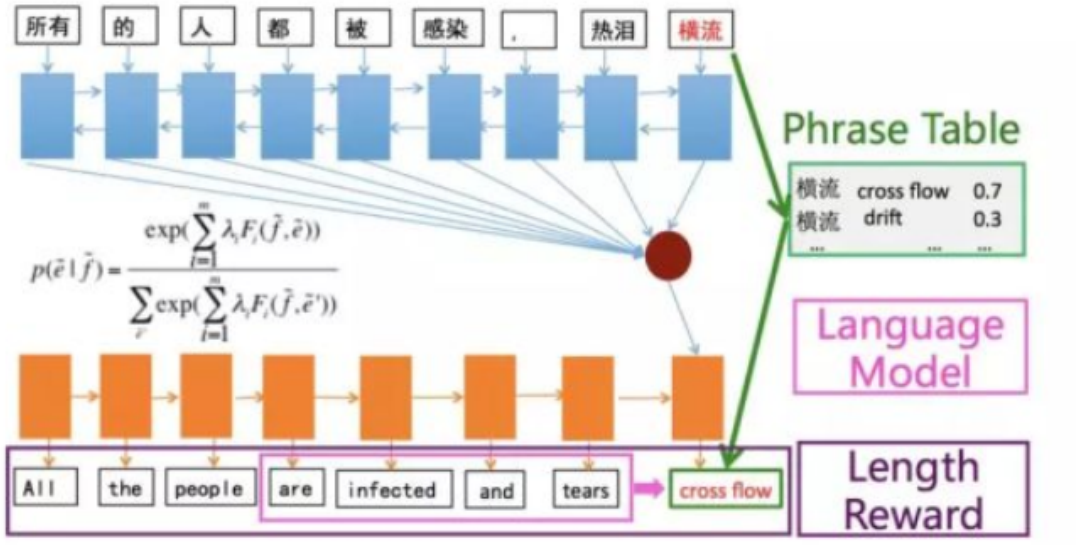


Target: Eventually , true , good , and beauty thoroughly defeated the ugly

"Addressing the Under-translation Problem from the Entropy Perspective " [zhao et al., To Appear in AAAI-19]

挑战二：数据稀疏

挑战三：引入知识



融合常识、世界知识

中巴 经贸 关系 得到 长足 发展

China-Pakistan
China-Palestine
China-Brazil
China-Bahamas

中巴 经贸 关系 在 金砖 框架 下 得到 长足 发展

Economic and trade relations have made great progress under the BRIC framework.

挑战四：可解释性

Source: 最终 真 善 美 彻底 打败 了 假 恶 丑

Black Box

Target: Eventually , true , good , and beauty thoroughly defeated the ugly

挑战五：语篇翻译

S1：我们加入霓虹，我们加入柔和的粉蜡色，我们使用新型材料。

S2：人们爱死这样的建筑了。

S3：我们不断的建造。

T1: We add neon and we add pastels and we use new materials.

T2: And you love it.

T3: And we cannot give you enough of it.