研究难点

仍有很多制约NLP发展的因素，这些因素构成了NLP的难点。而且要命的是，大多数是基础技术的难点。

**中文分词**，这条是专门针对中文说的。众所周知汉语博大精深，老外学汉语尚且虐心，更别提计算机了。同一个任务，同一个模型在英文语料的表现上一般要比中文语料好。无论是基于统计的还是基于深度学习的NLP方法，分词都是第一步。分词表现不好的话，后面的模型最多也只能尽力纠偏

**词义消歧：**很多单词不只有一个意思，但这个在今年BERT推出后应该不成问题，可以通过上下文学到不同的意思。另一个较难的是指代消歧，即句子中的指代词还原，如“小明受到了老师的表扬，他很高兴”，这个“他”是指“小明”还是指“老师”。

**二义性**：有些句子，往往有多种理解方式，其中以两种理解方式的最为常见，称二义性。

**OOV问题：**随着词嵌入技术大热后，使用预训练的词向量似乎成为了一个主流。但有个问题就是，数据中的词很可能不在预训练好的词表里面，此即OOV（out of vocabulary）。主流方法是要么当做UNK处理，要么生成随机向量或零向量处理，当然都存在一定的弊端。

**文本相似度计算：**是的你没有看错。文本相似度计算依旧算是难点之一。不过与其说难点，主要是至今没有一种方法能够从理论证明。主流认可的是用余弦相似度。但看论文就会发现，除了余弦相似度外，有人用欧式距离，有人用曼哈顿距离，有人直接向量內积，且效果还都不错。

**文本生成的评价指标：**文本生成的评价指标多用BLEU或者ROUGE，但尴尬的是，这两个指标都是基于n-gram的，也就是说会判断生成的句子与标签句子词粒度上的相似度。然而由于自然语言的特性（同一个意思可以有多种不同的表达），会出现生成的句子尽管被人为判定有意义，在BLEU或ROUGE上仍可能会得到很低的分数的情况。这两个指标用在机翻领域倒是没多大问题（本身就是机翻的评价指标），但用在文本摘要和对话生成就明显不合适了。