句子的最大匹配问题

陶晓鹏

2017.9.30

基于实例的机器翻译系统，预先存储一个语料库，其中每一项都是一个对译的双语句子对，通常也完成了句子中词的对齐。当输入一个待翻译句子，翻译步骤为三步：首先完成待翻译句子的分割，获得一系列已存在语料中的片段；然后以拷贝语料库已有译文的方式完成每个片段的翻译，得到一系列译文片段；最后将所有译文片段组装成整个句子的译文。其中第一步，我们称为匹配（match），匹配得到的单个片段越长，对应的译文越可靠，匹配得到的片段数量越少，组装阶段出错的机会就越少，因此，我们希望得到尽可能长、尽可能少的片段，我们把这个片段结果称为最大匹配（maxmatch）。

这个作业的目的就是设计和实现求解句子最大匹配的算法。例子如下，假设语料库为如下三个句子，

小明 今年 二十 岁
复旦 大学 的 学生 参加 了 歌唱 表演
他 是 交通 大学 的 老师

下面是一些输入和输出情况，

输入	小明 是 复旦 大学 的 学生
输出	[小明] [是] [复旦 大学 的 学生]
输入	他 今年 是 大学 的 学生 了
输出	[他] [今年] [是] [大学 的 学生] [了]