第二讲 双语数据预处理

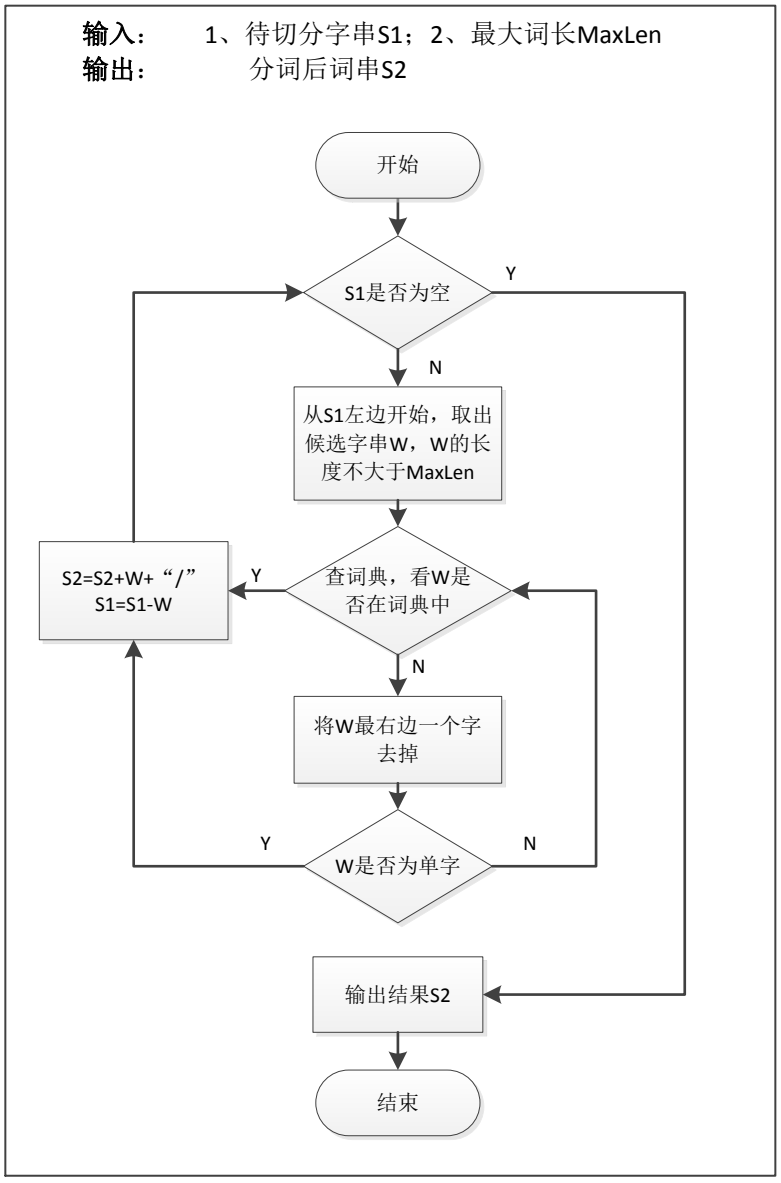
学习目标：了解和学习开发汉英双语数据预处理模块。

双语数据预处理是统计机器翻译系统构建的第一步，为词对齐处理提供分词后的双语数据。预处理的工作本质上就是双语数据的分词处理，与传统分词不同的一点在于需要对一些特定类型词汇进行泛化处理，如数字词汇“123.45”泛化为“$number”来代替原文。本讲中以汉英双语数据为处理内容。

本讲学习内容：

- **中文分词预处理**

采用传统基于词典的正向最大匹配法来完成中文分词。基本流程如图所示：



由于数字、日期、时间、网址等不可枚举，无法通过词典简单查找来分词。可以采用正则表达式或者自动机进行自动识别，并给予特殊名字进行泛化。例如：

数字类型	$number	如：123
日期类型	$date	如：1993 年 12 月 3 日
时间类型	$time	如：3:10
网址等类型	$literal	如：http://www.niutrans.com

实际上大家可以总结更多类型，并自行定义泛化名字进行替换原文。泛化的目的是为了有效解决数据稀疏问题。

需要注意一点的是，建议不要对组织机构名进行捆绑为一个词汇。例如将“东北大学信息学院”最好分成两个词“东北大学”“信息学院”。这样做的好处是为了有助于后面规则抽取模块抽取出更多翻译规则。

- **英文分词处理**

  相对于中文分词处理来说，英文分词主要处理三个问题：

  1) 将所有大写字母改为小写字母；
  2) 将英文句尾结束符与句尾最后一个单词用空格分开；
  3) 同样将数字、日期、时间、网址等不可枚举的类型进行识别，然后分别采用特殊名字进行泛化处理。

例如双语句对：

中文：4 月 14 日我买了 10 本书。

英文：I bought 10 books on April 14.

预处理结果：

中文：**$date** 我 买 了 **$number** 本 书 。

英文：i bought **$number** books on **$date** .

其它说明：

1) 中文的全角字符可以考虑改写为半角字符来处理；
2) 同一类型的泛化名字在中英文中最好一样，如中文/英文数字=>$number；
3) 也可以采用 CRF 或者语言模型来实现高性能中文分词；
4) 注意区分英文的句尾符号“.”和“Mr. Smith”的“.”；
5) 双语句对的泛化结果需要检查一致性，例如中文句子中包含$number，正常情况下，英文句子中也应该包含$number 等；
6) 目前有很多开源的分词工具可以被使用，如 NiuTrans 提供的双语数据预处理工具从 http://www.nlplab.com/NiuPlan/NiuTrans.YourData.html 下载。

本讲资源（UTF8 编码）：10 万行汉英双语句对和中文电子词典。