华为相似语料抽取方法

刘毅 <liuyi@torangetek.com>

2014 年 2 月 13 日

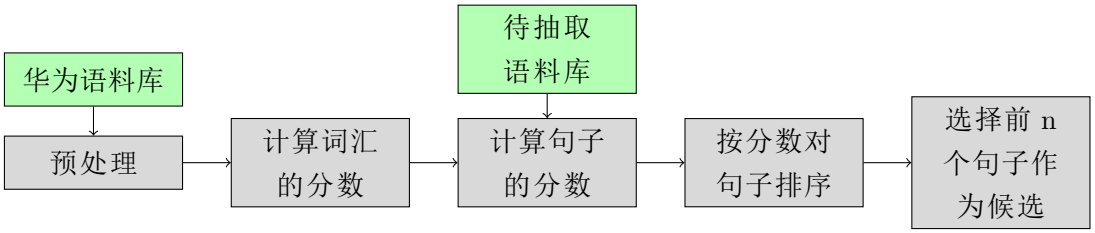
1 需求分析

项目需要在新闻语料库和专利语料库中抽取 500 万行左右的句子，要求抽取的句子与华为语料库中的句子相似，如使用了相同的词语，相同的语法，等特征。

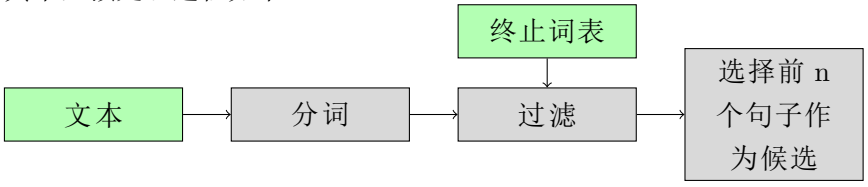
2 基本方法和原理

我们对新闻和专利语料库中的句子进行评分，然后根据分数进行排序，抽取分数较高者。

2.1 整体流程



其中，预处理过程如下：



2.2 分数的定义

我们对句子分数的定义有如下需求：

1. 被抽取的句子中的词在华为语料库中反复出现，频率较高。
2. 若源语料库中（新闻，专利）的某个句子中所有的词汇从未来华为语料库中出现过，那么该句子的得分为 0。若源语料库中的某个句子与华为语料库中的某句子完全相同，那么该句子应该得到相当高的分数。
3. 若源语料库中的某一句话的各个词汇分布在华为语料库中不同的句子中，该句子的分数也应该比较高。

2.3 词汇分数的定义

$$WS_i = \log(TF_{ij} \times \log DCN_i) \tag{1}$$

其中，TF_{ij} 指词频 (Term Frequency)，DCN_i 指文档覆盖数 (Documents Covering Number)，WS_i 表示该词汇 i 的分数 (Word Score), 分数越高表示该词在文档中越重要，越能代表这个文档的特征。

$$TF_{ij} = \frac{n_{ij}}{\sum_k n_{kj}} \tag{2}$$

其中，n_{ij} 是词 i 在文件 d_j 出现的次数，$\sum_k n_{kj}$ 是 d_j 中所有字词出现的次数之和。TF_{ij} 即词 i 在文件 d_j 中的频率。即在华为语料库中计算各个词出现的频率。

$$DCN_i = |\{j : t_i \in d_j\}| \tag{3}$$

其中，$|\{j : t_i \in d_j\}|$ 指包含词语 t_i 的文件数目。

2.4 句子分数的定义

$$SS = \sum_k WS_k \tag{4}$$

即句子中各个词汇的分数总和。

2.5 排序并抽取

最终对待抽取语料库中的句子根据分数进行排序，并取出 top n 即可完成。

3 代码说明

1. PassageP.py：PassageParse，对 JSON 格式的华为文本进行语法分析，提取出相应的文本信息。
2. GetWordRank.py：删除终止词并计算华为语料库的词汇分数。
3. GetExtractCorpus.py：计算句子分数，并且抽取句子。
4. Merge.py：融合双语抽取结果。
5. Split.py：分离成平行双语语料。